

University of Groningen

Topics in Corpus-Based Dutch Syntax

Beek, Leonoor Johanneke van der

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version

Publisher's PDF, also known as Version of record

Publication date:

2005

[Link to publication in University of Groningen/UMCG research database](#)

Citation for published version (APA):

Beek, L. J. V. D. (2005). *Topics in Corpus-Based Dutch Syntax*. s.n.

Copyright

Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

The publication may also be distributed here under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license. More information can be found on the University of Groningen website: <https://www.rug.nl/library/open-access/self-archiving-pure/taverne-amendment>.

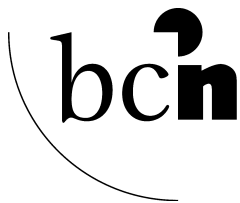
Take-down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): <http://www.rug.nl/research/portal>. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.

Topics in Corpus-Based Dutch Syntax

Leonoor Johanneke van der Beek



The work in this thesis has been carried out under the auspices of the Behavioral and Cognitive Neurosciences (BCN) research school, Groningen, and has been part of the Pionier project *Algorithms for Linguistic Processing* supported by grant number 220-70-001 from the Netherlands Organization for Scientific Research (NWO).



Groningen Dissertations in Linguistics 54

ISSN 0928-0030

Document prepared with L^AT_EX 2_ε

Printed by PrintPartners Ipskamp.

Rijks*universiteit* Groningen

Topics in Corpus-Based Dutch Syntax

Proefschrift

ter verkrijging van het doctoraat in de
Letteren
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
donderdag 10 november 2005
om 13.15 uur

door

Leonoor Johanneke van der Beek

geboren op 5 februari 1978
te Beuningen

Promotor: Prof. dr. ir. J. Nerbonne

Copromotores: Dr. G.J.M. van Noord
Dr. G. Bouma

Beoordelingscommissie: Prof. dr. J. Bresnan
Prof. dr. F. Van Eynde
Prof. dr. J. Hoeksema

Preface

I am indebted to many people for being able to finish this thesis. First of all my supervisor Gertjan van Noord. This dissertation has benefitted greatly from his comments on drafts and from the discussions we had, especially in the last months of the project. Moreover, I would like to thank him for allowing me to follow my own interests and make my own mistakes at all times.

I owe debt also to my promotor John Nerbonne and co-promotor Gosse Bouma for their valuable comments on my writings, and John also for his organizational and financial support as the head of the department. Joan Bresnan, Frank Van Eynde and Jack Hoeksema kindly agreed to be on my reading committee, for which I would like to thank them. This thesis benefitted particularly from the comments of Frank Van Eynde, who corrected mistakes, suggested improvements and offered additional data. I tried to incorporate these comments in this final version as well as possible, but obviously I remain responsible for all remaining flaws.

The seeds for almost all of this work were planted during my stay at Stanford University. I owe thanks to the CSLI people, Timothy Baldwin, John Beavers, Dan Flickinger, Stephan Oepen and Ivan Sag, for welcoming me in the LinGO Project and providing me with a stimulating work environment. I'm especially grateful to Timothy Baldwin. I learned a lot from our collaboration, the results of which form part of this thesis. Another highlight of my stay was Joan Bresnan's introduction to LFG. I certainly owe her my gratitude for discussion and support. I am thankful also to Rob Malouf for sending that one email that got me in Stanford in the first place.

Returning home was made easy by the many people who contributed to work and life in the Alfa-Informatica department. Wyke van der Meer, Anna Hausdorf and the secretaries ensured things ran smoothly in the department. Tanja Gaustad was always there to share everyday ups and downs with and Begoña Villada often initiated professional discussions or after-hour activities. Gerlof Bouma was always easily persuaded to share his linguistic intuitions or a coffee with me, and I thoroughly enjoyed our collaboration.

Stasinos Konstantopoulos made life easier by putting together the RuG thesis stylefile, and more agreeable by dragging me to *het Paard* every now and then. Eleonora Rossi helped me out with various practical issues, gave moral support when needed and makes a great “body of language” on the cover. Besides work, there were *aio*-dinners, movies and Gaioo/Grasp! meetings, pancake lunches, potluck dinners and (Friday) afternoon drinks—a big thank you to all who made life in Groningen so much fun.

Stasinos, Tanja and Begoña should be thanked once more for giving me the opportunity to practice the defense ceremony as a *paranimf*. I feel fortunate and proud that Irene Jansen and Eleonora Rossi agreed to stand by my side when it’s my turn, at last.

I would have never started this project if it weren’t for Víctor Sánchez-Valencia. His faith in me convinced me that I could do it and through our discussions on various linguistic topics I learned that I might very well love it. I wish I could still discuss linguistics, politics and life with you—or say thank you.

Finally, many thanks to my parents and my sisters. For always supporting me, for being there. *Dank je mam, dank je pap.*

Contents

1	Introduction	1
1.1	Corpus Linguistics	1
1.1.1	Data collection	1
1.1.2	Gradient patterns and probability	4
1.1.3	Resources for natural language processing	6
1.1.4	Evaluation	7
1.1.5	Corpus linguistics and this thesis	7
1.2	Methodological Preliminaries	9
1.2.1	Corpora	9
1.2.2	Tools	10
1.2.3	Statistics	11
1.3	Theoretical Framework	13
1.3.1	Lexical Functional Grammar	14
1.3.2	Optimality Theory	23
1.3.3	Stochastic OT	25
1.3.4	OT and LFG	26
1.4	Overview	29
2	Clefts	31
2.1	Introduction	31
2.2	Transitive Clefts	33
2.2.1	Differences between cleft clauses and other relative clauses	33
2.2.2	The c-focus	36
2.2.3	Agreement	37
2.2.4	The relative clause	44
2.2.5	Formalization	47
2.3	Intransitive Clefts	52
2.3.1	Differences between transitive and intransitive clefts . .	52
2.3.2	Differences between intransitive clefts and other complementizer constructions	56
2.3.3	The intransitive analysis	59

2.3.4	Formalization	62
2.4	Conclusion	64
3	Dative Alternations	67
3.1	Introduction	67
3.2	Previous Work	69
3.2.1	Linearization Constraints	70
3.2.2	The NP/PP alternation in English	73
3.3	Preliminaries	74
3.3.1	Resources and methodology	74
3.4	Linearization: the Double Object Construction	76
3.4.1	Pronominality	77
3.4.2	Gradient patterns	83
3.4.3	Weight	84
3.5	Linearization: Dative PP Shift	85
3.5.1	Weight	86
3.5.2	Pronominality and definiteness	88
3.6	The NP/PP Alternation	90
3.6.1	Lexical preferences	90
3.6.2	Weight and pronominality	92
3.6.3	Implementation in OT	94
3.7	More Factors in the Dative Construction?	99
3.8	Additional Evidence: the AcI	101
3.9	Conclusion	106
4	Determinerless PPs	109
4.1	Introduction	109
4.2	The Syntax of Determinerless PPs	110
4.2.1	Fixed determinerless PPs	111
4.2.2	Independent bare noun NPs	111
4.2.3	Compositional determinerless PPs	114
4.2.4	Prepositions selecting for determinerless NPs	117
4.2.5	Determinerless PPs as dependents	120
4.3	Extraction of PP-Ds	121
4.3.1	Introduction	121
4.3.2	Preliminaries	123
4.3.3	Prepositions selecting for determinerless NPs	125
4.3.4	Fixed determinerless PPs	126
4.3.5	Compositional determinerless PPs	128
4.3.6	Dependent determinerless PPs	131
4.4	Evaluation and Distribution of PP-Ds	132

4.5	Conclusion and Discussion	135
5	Countability	137
5.1	Introduction	137
5.2	Preliminaries	141
5.2.1	Countability classes	141
5.2.2	Lexical resources	144
5.2.3	Past research	145
5.3	Corpus-based Classification	147
5.3.1	Feature space	148
5.3.2	Methodology	151
5.3.3	Monolingual classifiers: design	152
5.3.4	Monolingual classifiers: results and discussion	155
5.3.5	Crosslingual classifiers	157
5.3.6	Crosslingual classification: results and discussion	159
5.3.7	Binary vs. three-way classifiers	162
5.3.8	Corpus-based approach: conclusion	162
5.4	Ontology-based Classification	163
5.4.1	Lexical resources for WordNet-based classification	165
5.4.2	Classifier design	166
5.4.3	Results and discussion	169
5.4.4	Ontology-based classification: conclusion	175
5.5	Conclusion	176
6	Conclusions and Future Work	179
6.1	Conclusions	179
6.2	Future work	182
	Appendix	185
	Bibliography	189
	Samenvatting	203
	Groningen Dissertations in Linguistics	209

Chapter 1

Introduction

In this introductory chapter, we start with a motivation for the corpus-based approach that we adopt in this thesis (section 1.1). We proceed in section 1.2 with some preliminaries about the corpora and statistics that are used, followed by a brief introduction in the theoretical frameworks assumed in section 1.3. At the end of the chapter, we give an outline of this thesis.

1.1 Corpus Linguistics

The chapters of this thesis each report an independent piece of research. The topics and the approach in each chapter vary widely, from purely theoretical linguistics in chapter 2 to machine learning in chapter 5. There is one thing that ties the chapters together: in each of the chapters, we make use of corpus data. The ways in which the corpus data is used vary just as widely as the the topics of the chapters. While corpora play only a supporting role as a natural source of examples in chapter 2, they form a crucial factor for estimating frequencies in chapter 3, identifying particular constructions in chapter 4 and training the classifiers in chapter 5.

1.1.1 Data collection

Even for traditional, theoretical linguistics research, corpus data form a valuable resource. (Electronic) text corpora may help theoretical linguists to find examples that support an analysis or counterexamples to some previously proposed analysis.

It is common practice to make up examples to illustrate a theoretical claim. This methodology allows the linguist to construct simple, short example sentences that abstract away as much as possible from anything that

is not relevant for the discussion.

For major, frequent constructions it is often very easy to construct such examples. For less probable constructions, however, it is often difficult to come up with a natural sounding example. Corpora may provide such examples for us. Although the sentences are often longer than made-up examples, they are more natural and easier to evaluate for grammaticality. As an extra advantage, corpus examples are usually extracted from large texts, so that the context, which may be important for evaluation and interpretation, is provided as well. Abney (1996) shows how grammaticality judgments may overlook perfectly grammatical though improbable parses, especially when presented out of context: although any subject will judge example (1) ungrammatical without the proper context, it is a fine sentence when uttered in the context of a map where large stretches of land are designated by capital letters and subdivided in pieces the size of one are and designated by lowercase letters. This problem is circumvented by the use of corpus data (in context).

(1) The a are of I.

Another advantage of corpus examples, is that they show the relevant construction in various different linguistic contexts. Browsing through these real world examples may reveal influencing factors that are not yet modeled in the analysis. As such, corpus data may drive further development of linguistic theory. In chapter 2, we almost exclusively use corpus examples to illustrate our analysis.

In some cases, the difficulty to come up with an example of a certain construction has lead to the conclusion that this construction must be impossible, i.e. ungrammatical. Corpus study may prove such claims wrong. Bresnan and Nikitina (2003) show that verb classes which were claimed not to participate in the dative alternation do in fact alternate. The alternative structures proved infrequent, rather than ungrammatical. Meurers (2004) discusses various claims in Germanic linguistics, which can be shown false based on corpus evidence. Counterexamples from corpora can also be found in chapter 2 of this book, where we falsify the claim that clefts have “regular” expletive subjects on the basis of corpus examples with demonstrative subjects, and in chapter 3, where we falsify the claim that only reduced object pronouns can shift (Zwart, 1996) on the basis of a number of counterexamples that we retrieved from corpus data.

The extraction of corpus examples usually proceeds in two steps: first the extraction of candidate examples and then the evaluation of the examples. The extraction of candidate examples is done on the basis of a search query.

Sometimes this query can be formulated in terms of the literal words in a sentence, but more often the query depends on more abstract notions such as part-of-speech or grammatical relation. In these cases, linguistic annotation is necessary. By adding meta-level linguistic information to the words in the corpus, linguists are able to search for abstract patterns, such as dative passives or it-cleft constructions. See Meurers (2004) for illustration of the process of retrieving examples via linguistic descriptions in the form of search queries.

In the evaluation step, the good examples are extracted from the total set of candidates. This set will also contain sentences which are grammatically fine, but do not contain the linguistic structure under investigation. For example, when looking for it-clefts in a corpus, one will come across sentences like example (2), where the relative clause is a modifier of the predicate. Annotation errors may produce more false candidates. On top of that, the corpus is likely to contain a number of ungrammatical sentences. The number of typos and grammar errors depends on the type of text, but even in heavily edited text one will come across ungrammaticalities: the examples in (3) are from a national newspaper. One can conclude that there is still an important role for grammaticality judgments in a corpus-based approach: separating the true positives from the false positives.

- (2) Het is een ontwikkeling die veel aandacht vraagt.
 it is a development that much attention asks
 It is a development that asks for much attention.
- (3) a. *Volgens de politie van Karlsruhe sloeg het voertuig
 according-to the police of Karlsruhe went the vehicle
 bij een inhaalmanoeuvre over de kop sloeg.
 during a take-over upside down went
- b. *Hij vind de teruggang van het aantal juristen in het
 He find_{1st} the decrease of the number lawyers in the
 parlement [...] zorgwekkend.
 parliament [...] disturbing

Linguists have often observed that the grammar of any natural language can generate an unbounded number of expressions. As a consequence, no corpus will ever contain all possible utterances of a language. This means that in case an appropriate search query does not yield any candidate examples, we cannot draw the conclusion that the construction is therefore ungrammatical. In order to draw any conclusion from the absence of a certain pattern in a corpus, one has to carefully evaluate the quality and size of the corpus, the frequencies of the individual construction components and the expected

frequency based on this information.

Summing up, corpora provide natural, real world examples for linguistic analyses which are easy to evaluate and which may reveal phenomena that were previously unaccounted for. As corpora also contain the infrequent or improbable expressions that one rarely comes up with while sitting at a desk grazing one's own intuitions, it forms a good test suite for existing assumptions about grammaticality and ungrammaticality. Finally, corpora may also provide some evidence that a particular pattern is not grammatical, although one must handle this evidence with care. As a final note, corpus data is an important source of linguistic evidence, but that is not to say it is the *only* appropriate type of evidence. See for example Wasow (2002, chapter 6) for a discussion of different types of linguistic evidence and how they can complement each other.

1.1.2 Gradient patterns and probability

Language is often modeled as categorical. Something is either a verb or a noun, either grammatical or ungrammatical, either productive or not. This is a useful simplification, which facilitated much linguistic research and has led to advanced symbolic models of language. Even in this thesis, strict categories are used to formulate generalizations. But it *is* a simplification. The idea that language is non-categorical is now dominating the field of computational linguistics. Jurafsky (2003) reported that 77% of the main conference (ACL) papers in 2000 built on stochastic models of language processing or learning. But also in other disciplines of linguistics we find research showing that various aspects of language are gradient. Ross was the first to show, for example, that syntactic categories are not discrete entities: “[L]et me propose that what is necessary is a relaxation of the claim that sequences of elements either are or are not members of some constituent class, like NPs, V, S etc. Rather, I suggest, we must allow membership to a degree” (Ross, 1973). This gradience is present in all areas of linguistics. Bod et al. (2003a) give an overview of probabilistic linguistic phenomena, ranging from language acquisition to language variation and from phonology to semantics.

Here, we will focus on gradient aspects of syntax and the use of quantitative corpus data for syntactic research. Abney (1996) already showed that many constraints on syntactically well-formed structures are in fact probabilistic in nature. For example, English plurals usually do not function as prenominal modifiers. Nevertheless, Abney lists a handful of examples from edited newspaper text, e.g. *the financial services industry, the bonds market*. In chapters 3, 4 and 5 of this book, we will come across various non-categorical phenomena: neither the ordering of objects in Dutch nor the selection for an

NP or PP recipient, a preposition's selection for a determinerless complement or a noun's countability are categorical.

Facing these gradient patterns, there are several strategies to choose from. One can choose to ignore the stochastic aspect and just allow for the most frequent pattern. As a consequence, all sentences containing one of the many less frequent patterns will be rejected. Alternatively, one can allow for all variants. This will lead to massive overgeneration and ambiguity. A third option is a probabilistic model of language, where different realizations are associated with different probabilities. Note that such a system does not exclude the usage of a symbolic grammar: one may simply attach weights to the rules of the symbolic grammar, or augment a symbolic grammar with a statistical disambiguation model, as in for example the Alpino grammar (Bouma et al., 2001, see also section 1.2). For an overview of ways in which probabilities may be incorporated in a grammar, see Manning (2003).

If one commits to research that allows non-categorical linguistic distinctions, corpora are indispensable. Introspective grammaticality judgments do not tell you which variant is most frequent, nor to what degree extra syntactic weight increases the chance of finding extraposition. In this thesis, we investigate various non-categorical phenomena. In chapter 3, we investigate which factors influence the ordering of the direct and the indirect object in Dutch. These factors are determined by making crucial use of quantitative data, such as difference in average weight in the shifted and the unshifted variant, or the frequency of a particular pronoun in each of the alternants. In chapter 4, we look at prepositions that combine with bare singular count nouns to form a determinerless PP, e.g. *per kind* 'per child', *richting huis* 'towards house'. Some of these prepositions combine almost always with determinerless complements, e.g. *per* in both Dutch and English, but others co-occur both with nouns with or without a determiner and some prepositions hardly ever form a determinerless PP. As a last example of gradual patterns discussed in this chapter, we try to determine the countability class of nouns automatically in chapter 5. Although a noun's countability appears to correlate rather straightforwardly with its potential to combine with certain determiners (*much* vs. *many*), we will see that 1) the notion countability in itself is not categorical and 2) the correlation between countability and various syntactic diagnostics for countability is not categorical. As we will see, it is nevertheless possible to classify nouns in countability classes based on their preferences for certain configurations.

Not all linguists have embraced quantitative data in linguistic research. Chomsky is strongly opposed to any notion of probability. "It must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term" (Chomsky, 1969, p.57). I will

not discuss all arguments and counterarguments in this discussion, as Abney (1996) and Manning (2003) convincingly argue for the use of probabilities in syntax and against the denial of probabilistic phenomena in language. I restrict myself to saying that none of the linguistic phenomena mentioned above and described in more detail in this thesis can be properly dealt with unless one accepts the existence of gradient patterns in linguistic theory and the use of quantitative data in accounting for these phenomena. Categorical distinctions simply prove empirically incorrect, while a treatment in terms of free variation fails to account for the influence of linguistic factors on the distribution of the alternants. Probabilistic data reveal a structure in language which is more fine grained than the notions ‘grammatical’, ‘ungrammatical’ and ‘free variation’. This structure builds on linguistic notions such as pronominality and definiteness. A theory of language may be expected to account for the influence of these factors on linguistic patterns, whether categorical or gradient.

1.1.3 Resources for natural language processing

Corpus data is indispensable in computational linguistics. Models of (various aspects of) language are trained on annotated or raw text corpora. For example consider preprocessors, which themselves make corpora more useful by tokenizing them, splitting them up in sentences, or enriching them with additional linguistic information such as part-of-speech or lemma. Although rule-based approaches to these tasks exist, most successful methods are supervised, crucially making use of corpus data for training.

But corpus data is also used to train models for more high level natural language processing: systems have been built to learn complex syntactic structures or even complete grammars from corpus examples (Bod et al., 2003b; Adriaans et al., 2004). In addition, symbolic systems may be complemented with a probabilistic component in order to model the gradient patterns mentioned earlier. The Alpino parser, for example, is based on a symbolic, handwritten grammar, but relies on corpus data for the training of its disambiguation module.

Corpora are also a natural source for other kinds of linguistic information, which may be extracted automatically. Lexicographers use large corpora to identify new words and automatic or semi-automatic methods have been developed to extract for example verb frames (Briscoe and Carroll, 1997; Keramides et al., 2004), semantically related words (Hearst, 1992; Lin, 1998; Finkelstein-Landau and Morin, 1999; van der Plas and Bouma, 2005), support verb constructions (Villada Moirón, 2004) or collocational prepositions (Bouma and Villada, 2002) from corpus data. In chapter 4 we will use cor-

pora to extract syntactically marked determinerless PPs automatically and in chapter 5 we use corpus data to classify nouns according to their countability class.

1.1.4 Evaluation

Another application of corpus data is evaluation. Manually annotated corpora may serve as a gold standard for natural language models or processors. Interesting as it is to discuss the beauty or the simplicity of a particular grammar of English, the value of a grammar is mainly determined by its capacity to correctly analyze all sentences of a language. This may be tested by applying the grammar to a set of real world sentences for which the correct analysis is known and count the number of mistakes the grammar makes. If the test set is representative, the performance on the test is a measure of the grammar's performance.

Furthermore, corpus data is used to monitor the effect of new additions to a system. As a model grows larger, it rapidly becomes impossible to oversee all consequences of a minor change. However, running the system over the test set will immediately tell you whether it improved or not. For illustration, see the increase in accuracy of the Alpino parser on the Alpino Treebank over a time span of 1000 days in figure 1.1. As the test set is used over and over again, performance on this set is not a representative measure of the absolute performance of the grammar, as it is likely that the data set will influence the work on the model. It should therefore only be used to obtain information about the performance relative to some other point in time.

In this book, we use corpus data for evaluation in chapter 4. By counting the number of determinerless PPs that we recognize as a particular syntactically marked construction, we measure the recall. In combination with an accuracy figure, this gives us an idea of the performance or quality of the system. Evaluation of the classification methods in chapter 5 is performed on an annotated test set of about a hundred hand-annotated nouns which were randomly extracted from a POS-tagged corpus (in addition to dictionary data and automatically classified data).

1.1.5 Corpus linguistics and this thesis

This book illustrates four applications of corpus data for syntactic research: corpora as a source of linguistic examples, gradient patterns in language, the automatic extraction of syntactically marked constructions from corpora and corpus-based classification. As such it formulates four answers to the question “why use corpora in syntactic research?”. In particular we show

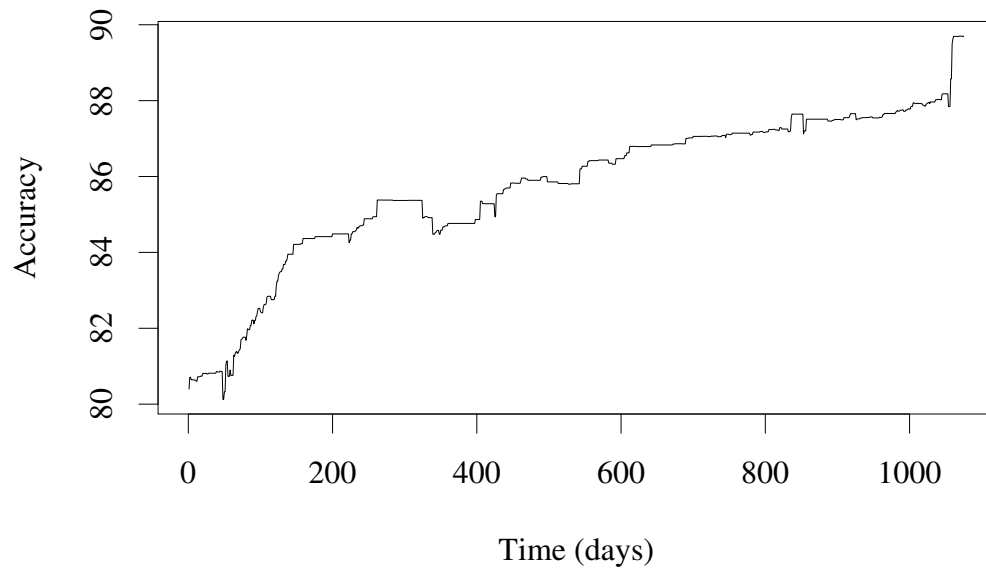


Figure 1.1: Accuracy of the Alpino parser on the Alpino Treebank.

that syntactically annotated corpora, manually built or machine-annotated, offer new opportunities for syntactic research.

The research project that led to this thesis was part of the Pionier program *Algorithms for Linguistic Processing*¹. This project aims at the development of a broad coverage parser for Dutch. The corpus-based research reported on in this thesis is meant to extend the coverage of the grammar and improve the performance of the parser. The Alpino parser in turn facilitates the development of large, syntactically annotated corpora, which make possible more research into the rules and regularities of Dutch, which will again improve the grammar of the parser. This book thus depends on and contributes to the availability of large, annotated corpora of Dutch.

1.2 Methodological Preliminaries

1.2.1 Corpora

Various corpora are used in this research. First of all, we used unannotated corpus data for finding relevant examples. We used three different national newspaper corpora: the 17M word **Volkskrant97** corpus and the equally large 1994 volumes of *NRC Handelsblad* and *Algemeen Dagblad*, which are both part of the **Twente Nieuws Corpus** (TwNC)². In case these did not provide the required examples, we used the **web** as our last resort.

For most tasks, we relied on annotated corpora. Two good sources of syntactically annotated material were available. First of all, the **Corpus of Spoken Dutch** (CGN), of which about 1M words are syntactically annotated. The annotation consists of dependency structures (Moortgat et al., 2001) and has been manually checked and corrected. The corpus contains spoken Flemish and Dutch of various genres, including (but not limited to) phone conversations, read aloud lectures and classroom discussions. A second manually annotated corpus that we used extensively is the **Alpino Dependency Treebank** (van der Beek et al., 2002a). This is a small size (145K words) corpus, consisting of the cdbl newspaper part of the Eindhoven corpus (den Boogaart, 1975). The sentences have been enriched with dependency structures as output by the Alpino parser (Bouma et al., 2001; van der Beek et al., 2002b, see also below) and familiar from CGN. The automatically generated annotations were all manually checked and corrected if necessary.

As these manually annotated corpora frequently proved too small for certain queries, we additionally made use of automatically annotated corpora.

¹<http://www.let.rug.nl/~vannoord/alp/>

²<http://wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html>

A total of 78M words of newspaper text from TwNC were available (the sections NRC1994, NRC1995, AD1994 and AD1995). Like the Alpino Treebank, the sentences were automatically annotated by the Alpino parser. But unlike the Alpino Treebank, no manual editing was performed. Despite the noise that may be introduced by parse errors, automatically annotated data are a rich source of linguistically relevant information, which one would not be able to extract without the annotation. This thesis illustrates this point amply. However, as the parser may systematically misanalyze a particular construction, one has to use the automatically annotated data with care. Before the data is used for a specific query, we investigate whether the parser makes systematic errors which influence the results. Only if no systematic bias is found, we use the automatically generated data. More information about the Alpino parser in the next section.

In addition to the corpus data, we also made use of **EuroWordNet**³ (Vossen and Bloksma, 1998) as a resource of linguistic knowledge. EuroWordNet is a manually composed hierarchical network of concepts, populated with words. The nodes are called synsets, as they contain sets of synonyms: words which express the same concept. The hierarchical relations express hyponymy and hypernymy relations between synsets. EuroWordNet furthermore connects the synsets of various languages, facilitating cross-linguistic generalizations.

1.2.2 Tools

Alpino Alpino is a wide-coverage grammar: it is designed to analyze sentences of unrestricted Dutch text. The grammar is based on the OVIS grammar (van Noord et al., 1999), that was used in the Dutch public transportation information system, but both lexicon and grammar have been extensively modified and extended. The lexicon contains about 100,000 entries at this moment (June, 2005) and more than 130 different verbal subcategorization frames are distinguished. Lexical information from the lexical databases Celex (Baayen et al., 1993), Parole and CGN (Groot, 2000) was used to construct the lexicon. Various unknown word heuristics further extend the lexical coverage of the system.

The grammar is rule based. The rules are written in a framework that is based on Head-Driven Phrase Structure Grammar (Pollard and Sag, 1994; Sag and Wasow, 1999). Following Sag (1997) construction specific rules are defined in terms of more general structures and principles. In total, the grammar contains over 330 rules.

³<http://www.illc.uva.nl/EuroWordNet/>

The parser usually generates a very large set of analyses for a sentence, upto 161,182 parses for the sentence in (4). A stochastic module selects the most probable parse from this set (Malouf and van Noord, 2004). The disambiguation model is based on a log linear maximum entropy model. In a nutshell, the model counts various features of the parses, e.g. dependency relations, the grammar rules that were used and unknown word heuristics. These features are associated with positive (preferred) or negative (dispreferred) weights. The ‘heaviest’ parse is then selected as the most probable one. The parser has an accuracy of about 85.5%, measured over the dependency relations.

- (4) Aling maakte deel uit van een kopgroep van zeven renners die bijna een ronde voorsprong had op het peloton.
 Aling made part out of a breakaway of seven cyclists that almost a lap lead had on the pack
Aling was part of a breakaway of seven cyclists that was almost one lap ahead of the pack.

Dt_search The advantage of syntactically annotated corpora is that one can formulate queries that refer to syntactic rather than lexical units. One can search for ‘it’-subjects with plural verb agreement or pronominal indirect objects. The search tool **dt_search** (Bouma and Kloosterman, 2002) facilitates the formulation of these and more complex queries. It is built on top of the more general XML search tool XPath. **Dt_search** queries may refer to POS, dependency relation, word form and string position and. It is thus possible to search for hierarchical relations, such as the verb heading a determinerless PP, or a dative passive (i.e. a noun phrase which is identified both as a subject and as an indirect object).

1.2.3 Statistics

On various occasions, simple statistic measures are used, for instance to quantify the association between two variables. We will briefly discuss the relevant measures.

Entropy Entropy is a measure of uncertainty or unpredictability. A low entropy indicates that there is very little uncertainty about the value of a variable. We apply entropy in the detection of syntactically marked determinerless PPs. For instance, we set a minimum PP-verb entropy, meaning that if we look at the verb that heads a determinerless PP, the entropy must be higher than some threshold value. By setting this minimum, we ensure

that the PP combines with various verbs and effectively exclude fixed PP complements. The formula for the entropy H of a variable X is a weighed average of the probabilities of all possible outcomes, expressed in bits (hence the negative logarithm to the base 2):

$$H(X) = - \sum_{x \in X} p(x) \log_2 p(x)$$

Entropy Reduction By adding knowledge to a system, one reduces the uncertainty. The information gain can be quantified by comparing the total entropies of the original system and the final system. This measure is applied to quantify the influence of the verb lexeme and the syntactic category of the direct object on the dative alternations in Dutch.

$$H_{1-2} = |H_1 - H_2|$$

Pearson’s Chi Square Test Chi square (χ^2) is a non-parametric test of statistical significance that is applied to contingency tables. It tells you the degree of confidence you can have in rejecting the null hypothesis (i.e. the hypothesis that the distribution is due to chance). In other words: it tells you with some degree of confidence whether or not two variables are independent. The test is applied in chapter 3 to test whether realization of the recipient argument as an NP or a PP is independent of the verb lexeme. χ^2 compares for all cells in the contingency table the observed frequencies (O) with the expected frequencies (E). If this difference is large, we can reject the null hypothesis of independence. The advantage of χ^2 is that it does not assume that probabilities are distributed normally. The disadvantage is that the test is unreliable with small numbers: it should not be used if the total sample size is less than 20 or if the total sample size is between 20 and 40 and the expected value of any cell is 5 or below.

$$\chi^2 = \sum_{i,j} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

Log-Likelihood Ratio The log-likelihood ratio (G^2) is a measure of association, similar to the Pearson’s χ^2 test. However, it is less sensible to low-frequency data. Bouma and Villada (2002) showed that log-likelihood outperformed χ^2 on the task of collocational PP extraction. We thus opt for log-likelihood on the related task of extracting syntactically marked determinerless PPs. For two-way contingency tables, the statistic simplifies to the formula

$$G^2 = 2 * \sum_{i,j} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

This ratio is called the *likelihood-ratio chi-squared statistic* (Agresti, 2002).

1.3 Theoretical Framework

Four topics in Dutch syntax are discussed in this thesis: the it-cleft construction, the dative alternation, determinerless PPs and noun countability. We formalize the regularities of these constructions in rules and constraints and we identify the relevant lexical items and features. This linguistic information should be cast in some theoretical framework. The diversity of the topics puts heavy constraints on this framework: it must allow for two constituents mapping to a common syntactic role (in order to account for the Dutch it-clefts), it must allow for gradient patterns and non-categorical distinctions (in order to account for the various linguistic factors influencing the dative alternation) and it should allow for an information rich lexicon (in order to store the determinerless PP and noun countability information).

The basic syntactic framework assumed in this thesis is Lexical Functional Grammar (LFG). This framework separates constituent structure from functional structure, which facilitates the account of constructions where the two levels of syntax do not coincide, as we will see is the case for Dutch it-clefts. It is furthermore a lexicalist theory of syntax. This is crucial for our account of determinerless PPs. As we will see in chapter 4, the lexical specifications of the preposition or the noun determine in which type of determinerless PP the word can participate and what type of modification is allowed. Naturally, the lexicon is also the place to store the results of our noun countability classification efforts in chapter 5.

But classic LFG is a symbolic grammar type which does not allow for gradient patterns. Therefore it does not allow us to express the non-categorical distributional differences that we find in the dative alternation: certain linguistic factors favor realization of the recipient argument as a PP, while other factors prefer the double object construction realization. Optimality Theory (OT) does allow for the modeling of (large numbers of) violable constraints on a particular construction. The stochastic extensions to OT furthermore facilitate the modeling of distributional patterns in linguistic variation, without losing the linguistic insights. (Stochastic) OT would thus allow us to model our account of the Dutch dative alternation in chapter 3. Fortunately, LFG and OT may very well be combined. Various modes of

combination have been discussed in the literature. These range from variants in which the LFG component is minimal and all the explanatory power is put in the OT component (Bresnan, 2000, 1999, 2002, 2001a; Smolensky and Legendre, 2005), to full-fledged LFG grammars with OT modules to account for certain preferences (Frank et al., 2001). As we rely on a lexicalist type of grammar, we envisage a model similar to the one described in Frank et al. (2001) and briefly discussed below.

That being said, the results obtained in the following chapters do not rely heavily on any particular theoretical framework. The chapters offer linguistic insight and this information is poured in a framework which allows us to express the relevant notions. But any grammar in any theoretical school will have to account for the agreement facts of Dutch *it*-clefts and the linguistic factors influencing the dative alternation. And any grammar will need to have access to information about which prepositions and which nouns form determinerless PPs and the countability of nouns. We will come back to this point at the end of section 1.3.4.

In the remainder of this section we will first give a short introduction in LFG, followed by introductions in classic OT, stochastic OT and OT-LFG.

1.3.1 Lexical Functional Grammar

Various good introductions in Lexical Functional Grammar (LFG) have been published in the last couple of years: Bresnan (2001b), Dalrymple (2001), Falk (2001). This introductory section on LFG is not meant to be an alternative to these excellent books; any reader interested in learning LFG, is referred to those works. We included this introductory section in this thesis so that any linguist, regardless his or her theoretical background or preferred school, can read, understand, and evaluate the linguistic analyses in chapter 2 and 4. We therefore introduce the basic LFG terms and concepts below. The topics below are all uncontroversial and generally accepted parts of ‘classical’ LFG. Novel features or extensions to the theory are discussed in the chapters that follow.

F-structure

LFG separates two levels of syntax: the surface form or constituent structure (c-structure) and the functional structure (f-structure). C-structure represents the overt linear and hierarchical organization of words into phrases, while f-structure represents the more abstract, functional organization of the sentence in the form of grammatical functions such as subject and object.

These two structures are parallel: one is not derived from the other, but both are built in parallel.

The functional information encoded in f-structure takes the form of sets of attribute-value pairs, usually depicted in square brackets. Examples of attributes are PRED for predicate, SUBJ for subject and TENSE for verb tense. A large part of the f-structure attributes encode grammatical relations. These include the argument function SUBJ (subject), OBJ1 (direct object), OBJ2 (indirect object), OBL (oblique), COMPL (verbal complement) and XCOMP (open complement).⁴ In addition, there are non-argument functions, such as TOP (topic), FOC (focus) and ADJ (adjunct).

Each attribute has a unique value (**uniqueness condition**). That is, attributes may not have multiple values (but as we will see, it is possible for multiple attributes to have the same value). These values can be atoms, semantic forms, (embedded) f-structures or sets of f-structures. We give an example of each of these possibilities. A sample atomic value is ‘past’, which is an appropriate value for the attribute TENSE. There is only one attribute which takes a semantic form as its value, and that is the attribute PRED. Semantic forms may be simplex, or complex. A complex semantic form encodes the arguments it selects for. Simplex predicates do not select any arguments. An example of a simplex predicate is ‘Bo’, an example of a complex predicate is ‘sleep<SUBJ>’. As mentioned, f-structures may themselves serve as values. An example of a value which is in turn an f-structure can be found in example (6). The value of the attribute SUBJ is the f-structure in (5).

- (5) *Bo*
- $$\left[\begin{array}{ll} \text{PRED} & \text{'Bo'} \\ \text{NUM} & \text{sg} \end{array} \right]$$
- (6) *Bo slept*
- $$\left[\begin{array}{ll} \text{PRED} & \text{'sleep<SUBJ>'} \\ \text{TENSE} & \text{past} \\ \text{SUBJ} & \left[\begin{array}{ll} \text{PRED} & \text{'Bo'} \\ \text{NUM} & \text{sg} \end{array} \right] \end{array} \right]$$

Finally, we have values consisting of sets of f-structures. These are the values of the attribute ADJ, or adjunct⁵. As noted by Kaplan and Bresnan (1982),

⁴Our use of OBJ1, OBJ2 and OBL is partly a notational variant on most work in LFG and partly a simplification. In classical LFG, the primary object is labeled OBJ. In addition, OBJ_θ and OBL_θ are used to indicate families of relations, where the *θ* subscript refers to the semantic role associated with the argument, i.e. source, goal or theme.

⁵There are other attributes which take set values, e.g. conjunctions, but we will not

modifiers, but not arguments, can be multiply specified in a sentence. All specifications are encoded under one attribute in the f-structure as members of a set. For example, in sentence (7), we have two sentential modifiers, *well* and *yesterday*, which form the two members of the set which is the value of ADJ.

- (7) *Bo slept well yesterday*
- $$\left[\begin{array}{ll} \text{PRED} & \text{'sleep<SUBJ>'} \\ \text{TENSE} & \text{past} \\ \text{SUBJ} & \left[\begin{array}{ll} \text{PRED} & \text{'Bo'} \\ \text{NUM} & \text{sg} \end{array} \right] \\ \text{ADJ} & \{ \left[\begin{array}{ll} \text{PRED} & \text{'well'} \end{array} \right], \left[\begin{array}{ll} \text{PRED} & \text{'yesterday'} \end{array} \right] \} \end{array} \right]$$

Sometimes, two attributes have the same value. If this value is atomic, it is simply repeated, e.g. if both the subject and the object of a sentence are singular. Things are trickier when the value is a semantic form. Since each instance of a semantic form is unique, a duplication of the semantic form should be thought of as having a unique index, indicating that the two are distinct objects. As f-structures always have a predicate, it is not good practice to repeat an f-structure if it is shared between two attributes. Instead, what we use is the notation in example (8), where the line indicates that the structure on one end of the line is shared with the attribute on the other end of the line.

- (8) *Bo seemed to sleep*
- $$\left[\begin{array}{ll} \text{PRED} & \text{'seem<XCOMP>'} \\ \text{TENSE} & \text{past} \\ \text{SUBJ} & \left[\begin{array}{ll} \text{PRED} & \text{'Bo'} \\ \text{NUM} & \text{sg} \end{array} \right] \\ \text{XCOMP} & \left[\begin{array}{ll} \text{PRED} & \text{'sleep<SUBJ>'} \\ \text{SUBJ} & \text{---} \end{array} \right] \end{array} \right]$$
-

All f-structures must meet two wellformedness conditions in addition to the uniqueness condition mentioned above: completeness and coherence. **Completeness** tells us that all arguments that are specified in a complex PRED value must be realized.⁶ It follows from this requirement that a sentence like:

- (9) *Bo seemed

discuss these here.

⁶For a formal definition, see Kaplan and Bresnan (1982)

is ungrammatical, as we just saw that the predicate of *to seem* introduces an argument XCOMP. This argument must be realized in order to meet the completeness condition.

The **coherence** condition furthermore states that argument functions must be governed by the semantic form of the PRED of the f-structure it appears in. That is, we cannot add extra arguments to an f-structure. Thus, example (10) is ungrammatical, as the predicate ‘sleep<SUBJ>’ only introduces a subject and no object.

(10) *Bo slept the bed

C-structure

The linear and hierarchical organization of words into phrases is modeled in c-structure, usually depicted as a tree. The leaves of the tree are occupied by words. This means that words are the smallest units of a c-structure, and that a c-structure rule cannot refer to any unit smaller than a word. In other words, LFG adheres to the principle of **lexical integrity** as formulated in Lapointe (1980):

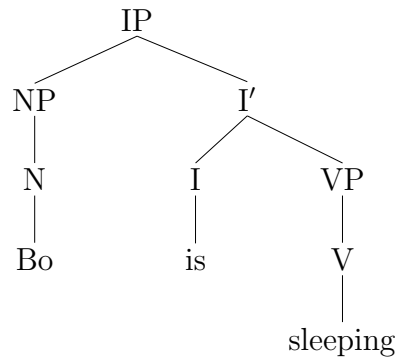
(11) No syntactic rule can refer to elements of morphological structure (Lapointe, 1980, p. 8)

The LFG c-structure assumes four major lexical categories: N(oun), V(erb), A(djective) and Adv(erb). These project the corresponding phrases NP, VP etc. In addition to lexical categories, there are also functional categories such as I and C. In Dutch, I is the category for tensed verbs in main clauses, assuring verb second, and C is used for complementizers. Again, they project the corresponding phrases IP and CP. The difference between lexical and functional categories will be discussed in the next section.

The way in which the categories are organized obeys the general rules of *X-bar theory* (Jackendoff, 1977; Chomsky, 1986). Lexical and functional categories head the phrases they project and combine with complements and specifiers. Two levels of projection are assumed: X' and XP. As the f-structure wellformedness conditions already restrict the space of possible structures, few additional, general restrictions on c-structure exist in LFG. For example, LFG does not assume binary branching: a node may have any number of daughter nodes. Furthermore, any category is optional, even lexical heads. Does this mean that headless constructions are allowed, violating the principle of endocentricity? Yes, although some weaker notion of endocentricity still applies to c-structure (Bresnan, 2001b, p. 133), and the f-structure wellformedness conditions ensure that every phrase is headed on

the level of f-structure. Thus, VPs are allowed in Dutch main clauses, even if the only verb is realized in I, and without the need to assume a trace or slash. Although the absence of a strict headedness condition would facilitate the insertion of extra syntactic categories, **economy of expression** tells us that syntactic structure nodes are not used unless required by independent principles such as completeness, coherence or semantic expressivity (Bresnan, 2001b, p. 91), thus preventing the spurious ambiguities that the insertion of void syntactic nodes would give rise to. A sample c-structure is given in (12).

(12) *Bo is sleeping*



Mapping from c-structure to f-structure

C-structure and f-structure are two aspects of one and the same linguistic object. Each node in a c-structure corresponds to an f-structure, as illustrated in figure 1.2. This section describes the way in which this correspondence is established.

The mapping between c-structure and f-structure is defined through f-structure annotations on c-structure nodes. The terminals of the syntactic tree are lexical items, which are lexically specified with functional constraints. These constraints are passed on to the preterminals and syntactic nodes following general (X-bar based) schemata and c-structure rules with functional annotations. We describe each of the three components (lexical entries, general schemata and c-structure rules) below.

Example (13) shows the lexical entry for the word *Bo*. It specifies the syntactic category of the word to be N. In addition, it constrains the f-structure associated with its mother node, indicated by the \uparrow , to have an attribute PRED and an attribute NUM with the values ‘Bo’ and ‘sg’ respectively.

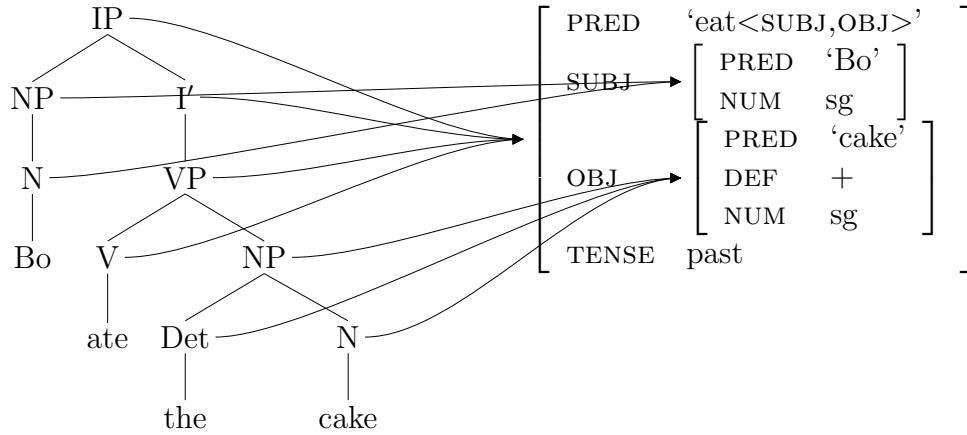


Figure 1.2: c/f-structure correspondence.

$$(13) \quad \begin{array}{lll} Bo & N & (\uparrow \text{PRED}) = \text{'Bo'} \\ & & (\uparrow \text{NUM}) = \text{sg} \end{array}$$

Lexical entries may also refer to grammatical functions. The lexical entry for the verb *am* in example (14) selects for a subject and an open complement and furthermore requires its subject to be first person singular.

$$(14) \quad \begin{array}{lll} am & V & (\uparrow \text{PRED}) = \text{'be<SUBJ,XCOMP>'} \\ & & (\uparrow \text{SUBJ PERS}) = 1 \\ & & (\uparrow \text{SUBJ NUM}) = \text{sg} \end{array}$$

The \uparrow symbols refer to the f-structures associated with the preterminals. If the preterminal is heading a phrase, it will pass on all functional constraints to the mother node. This follows directly from X-bar theory and the notion of headedness. In LFG this is indicated with the annotation $(\uparrow=\downarrow)$ on the head daughter, saying that all information on the current node (\downarrow) is shared with the mother node (\uparrow). Similar annotation schemata exist for other kinds of nodes. A non-head daughter in a lexical projection is annotated $(\uparrow \text{CF})=\downarrow$, where CF stands for Complement Function, and $\text{CF}=\{\text{OBJ1}|\text{OBJ2}|\text{OBL}\}$. An example of a non-projecting daughter in a lexical projection is the object NP in figure 1.2. All information on this node will project to a complement function (in this case direct object) of the mother node's f-structure. In functional projections, two daughters project all their information up to the mother: the head and the functional co-head. Examples are the I-node (head) and the VP-node (co-head) in example (12). Non-projecting nodes in functional projections are associated with discourse functions (DF). The annotation is $(\uparrow \text{DF})=\downarrow$, with $(\text{DF}=\{\text{TOP}|\text{FOC}|\text{SUBJ}\})$. Phrasal nodes may

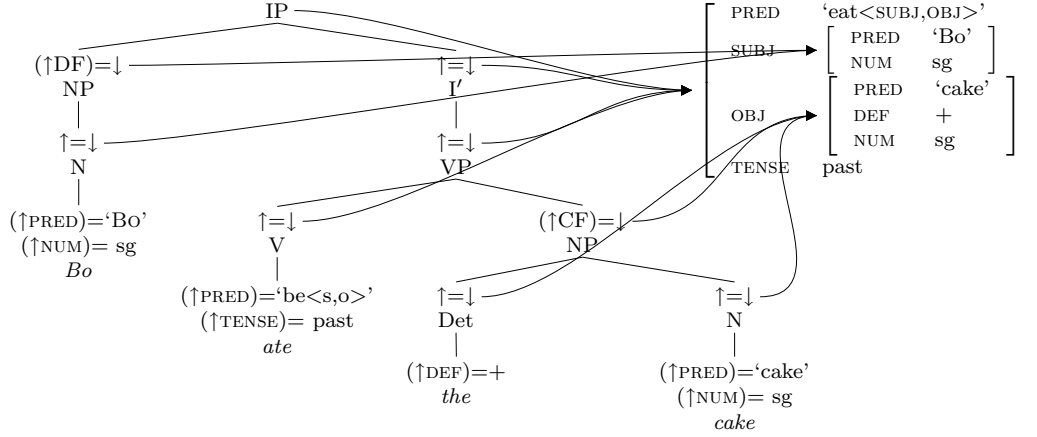


Figure 1.3: c/f-structure correspondence: lexical and predictable annotations.

have non-projecting sisters. These fill the non-argument functions (NAF) TOPIC, FOCUS and ADJ. The annotation on the node: $(\uparrow NAF) = \downarrow$. So the topic and focus functions may be realized in both lexical and functional projections. Together, these rules give us all predictable annotations. A summary:

- (15) a. A lexical head of a phrase is annotated $\uparrow = \downarrow$
- b. A functional head is annotated $\uparrow = \downarrow$
- c. A functional co-head is annotated $\uparrow = \downarrow$
- d. A non-head daughter in a lexical projection is annotated $(\uparrow CF) = \downarrow$ ($CF = \{OBJ1|OBJ2|OBL\}$)
- e. A non-head daughter in a functional projection is annotated $(\uparrow DF) = \downarrow$ ($DF = \{TOP|FOC|SUBJ\}$)
- f. A non-projecting sister of a phrasal node may be annotated $(\uparrow NAF) = \downarrow$ ($NAF = \{TOP|FOC|ADJ\}$)

For illustration, we decorated the nodes in the tree from fig. 1.2 with their functional annotations, as shown in fig. 1.3. Note that although the schemata in (15) leave open whether *Bo* is the subject, topic or focus and whether *the cake* is a direct object, an indirect object or an oblique, Coherence and Completeness rule out all other options except the correct one (SUBJ and OBJ1 respectively).

In addition to the predictable annotations, there are unpredictable and language specific annotations, which must be explicitly specified in c-structure rules. For example, in Dutch, a noun indicating some kind of measure may combine with an N', which indicates the substance. The measure noun acts

as the syntactic head of the construction, which is modified by the substance noun following it. The results is something similar to the N_1 of N_2 construction in English (*a pound of sugar*, *a bunch of flowers*). To account for this construction, we might write a c-structure rule as in (16).

$$(16) \quad N' \Rightarrow \begin{array}{ccc} & N & N' \\ & \uparrow=\downarrow & \downarrow \in (\uparrow \text{ADJ}) \\ & (\downarrow \text{NFORM})=\text{meas} & \end{array}$$

Recall that ADJ is a set valued attribute. The annotation $\downarrow \in (\uparrow \text{ADJ})$ defines the f-structure corresponding to the N' -node to be a member of the set of f-structures which is the value of the ADJ. The functional annotation ' $(\downarrow \text{NFORM})=\text{meas}$ ' defines the head noun to be of noun type 'meas', i.e. a measure noun. LFG is based on **unification**. This means that parsing (or generating) will succeed unless the noun is otherwise specified for the feature NFORM. If the head noun simply does not have a feature NFORM, it will receive the attribute value pair 'NFORM=meas' as a result of the annotation. We therefore call this type of annotation **defining equations**. If the grammar writer wishes to restrict the application of the rule in (16) to nouns which are predefined as measure nouns, this is possible by using a **constraining equation** instead, indicated by an underscore c . The relevant equation would be ' $(\downarrow \text{NFORM})=c\text{meas}$ '.

Several other options are available for constraining the application of a rule or to add functional information to c-structure nodes. It is possible to formulate a negative constraint, e.g. $\neg(\uparrow \text{TENSE})$ for a non-tensed phrase, or $(\downarrow \text{NFORM}) \neq c\text{meas}$ for a nominal which is not a measure noun.

Existential constraints simply require that the f-structure corresponding to a node has a particular attribute, without constraining its value. An annotation $(\uparrow \text{TENSE})$ thus requires the dominating node to be tensed.

Optional constraints, printed in parentheses, may or may not be instantiated. They may be used, for instance, to account for clitic doubling: if the clitic is used independently, its PRED value must be 'pro', similar to regular pronouns. However, if it is doubled, it cannot have any predicate, because the NP that it doubles already has a predicate and the two would fail to unify. The clitic is then annotated $((\uparrow \text{PRED})=\text{'pro'})$. Lexical entries with optional constraints can be viewed as shorthand for the combination of two entries: one with the relevant constraint, and the other one without.

Furthermore, it is possible to combine constraints in a conjunction or a disjunction. Example (17) is just a silly way of saying $(\uparrow \text{PERS})=c3$, and example (18) is an alternative to $(\uparrow \text{PERS}) \neq c3$.

$$(17) \quad (\uparrow \text{PERS}) \neq c1 \wedge (\uparrow \text{PERS}) \neq c2$$

$$(18) \quad (\uparrow \text{PERS}) =_c 1 \vee (\uparrow \text{PERS}) =_c 2$$

Finally, it is important to know that it is possible to refer to a large number of paths through an f-structure. The technique that makes this possible is **functional uncertainty**. The technique was introduced by Kaplan et al. (1987). Kaplan and Zaenen (1989) applied it in their treatment of long distance dependencies. We have already seen a simple example of it above: when summarizing the X-bar based annotation schemata in (15), we used abbreviations like CF, DF and NAF for complements functions, discourse functions and non-argument function. Their denotation is a set of grammatical functions, and any member of that set will successfully instantiate the condition. In other words: the exact function was left uncertain. Building on this, it is possible to describe longer paths through an f-structure. Example (19) for example states that the focus of the dominating node fills a grammatical function that is arbitrarily deep embedded in open or verbal complements of this node (describing for example the path between a wh-word and its ‘trace’). The star (\star) is the Kleene star, indicating *zero or more* of the preceding element. In this case, the preceding element is either an open complement or ($|$) a verbal complement. GF stands for $\{\text{SUBJ}|\text{OBJ1}|\text{OBJ2}|\text{OBL}|\text{COMP}|\text{XCOMP}|\text{ADJ}\}$. Correct instantiations of the complete path would be (COMP OBJ), (COMP COMP ADJ), or (COMP XCOMP XCOMP OBL), for example.

$$(19) \quad (\uparrow \text{FOC}) = (\uparrow \{\text{XCOMP}|\text{COMP}\} \star \text{GF})$$

Not only can we refer to f-structures that are arbitrarily deep embedded, we can also refer to f-structures that enclose the reference point (i.e. in which the reference point is embedded). The technique for this is called **inside-out functional uncertainty**. Again, we explain the technique with an example. In (20), it is stated that the f-structure whose locative oblique is the projection of the annotated node has ergative case.

$$(20) \quad ((\text{OBL}_{loc} \uparrow) \text{CASE}) = \text{erg}$$

Besides c-structure and f-structure, various other levels of syntax have been proposed in the LFG literature, most notably a(rgument)-structure. We will not discuss these here, as they are not relevant to the topics discussed in this book. However, it is clear that mappings between these other structures may take the same form as the f-structure annotations on c-structure nodes that we saw above.

1.3.2 Optimality Theory

From LFG syntax, a new framework has emerged, namely OT-LFG, or Optimality Theoretic LFG. Before we discuss this particular incarnation of OT syntax in the next section, we first describe the basics of ‘classic’ OT, as it was developed in the field of phonology in the early nineties (Prince and Smolensky, 1993).

In classic OT, the task of generating or interpreting a linguistic object is split up in two subtasks. First a set of candidate realizations or interpretations is generated, and from these output candidates the most optimal is picked in a second step.

The generator component is called GEN. It is a function from the input to the set of output candidates. In phonology, the input consists of an underlying form for generation and a surface form for interpretation. It is assumed that the input is unrestricted. That is, no linguistic explanation may depend on the assumption that some underlying form does not exist in some language. The generator function GEN is furthermore assumed to be universal. Combined with the unrestricted input, this results in a universal set of possible output candidates. This principle is called Richness of the Base (Prince and Smolensky, 1993).

Evaluation of the candidates is based on a ranked set of constraints. The candidate which best meets the constraints, is the optimal (or grammatical) candidate, even if it violates some constraints. The constraints are strictly ranked with respect to each other. If constraint C_1 dominates constraint C_2 ($C_1 \gg C_2$), one violation of C_1 is worse than any number of violations of constraint C_2 . Two constraint families are distinguished: markedness constraints and faithfulness constraints. The first requires the output to conform to certain structural patterns. The faithfulness constraints require that the output contains all information from the input and no more. Clearly, the two sets of constraints potentially contradict each other: a markedness constraint may require clauses to have subjects, but if the verb does not specify a subject role, as in the English *rain* and the Dutch *regenen*, this constraint can only be met by introducing a subject which was not in the input, thus violating faithfulness. The set of constraints CON is assumed to be universal. As the input was unrestricted and GEN was also universal, the only source of cross-linguistic variation is the ranking of the constraints, which determines their importance in evaluation.

An OT analysis is usually depicted in a table called tableau, as in fig. 1.4. The tableau only visualizes the evaluation step, as that is the component where all the explanatory power of the model lies. In the upper left cell, we put the input. The relevant constraints are in the upper row, from left to

[INPUT]	C ₁	C ₂	C ₃	C ₄
☞ Candidate 1			*	*
Candidate 2		*!		**
Candidate 3		*!		
Candidate 4	*!	*		*

Figure 1.4: Sample OT tableau.

[INPUT]	C ₁	C ₂	C ₃	C ₄
☞ Candidate 1			*	*
Candidate 2		*		**
☞ Candidate 3		*		*
Candidate 4	*!	*		

Figure 1.5: Sample OT tableau with constraint stratum.

right in descending order of importance. In the left column, we list the relevant output candidates. Constraints and candidates which are not relevant to the analysis are left out for clarity and simplicity. Note that one must guarantee that all candidates that are not depicted in the tableau are ruled out somehow. Each cell of the tableau then indicates whether or not a particular candidate violates a constraint. An empty cell indicates that there is no violation, a * indicates a violation and a fatal violation is indicated with a !*. Grammatical output candidates are marked ☞.

Some constraints are equally strong. In this case, they are unranked relative to each other. We call a set of unranked constraints a **stratum**. In a tableau, the constraints in a stratum are not separated by vertical lines. By using strata, it is possible to have two candidates which have the same violation profile. In this case, both candidates are grammatical, and free variation is expected (with both candidates surfacing in about 50% of the cases). Figure 1.5 illustrates optimization with a stratum and two winners.

Constraints may ‘gang up’ to dominate other constraints. This is called **constraint conjunction**. The conjunction is always higher ranked than all its component constraints. For example, assume we have the following constraint ranking:

$$(21) \quad C_1 \gg C_2 \gg C_3$$

The conjunction of C_2 and C_3 ($C_2 \& C_3$) may outrank the first constraint (22). Constraint conjunction is the only way of modeling cumulativity effects

(multiple violations of lower ranked constraints outweighing one violation of a higher ranked constraint) in classic OT.

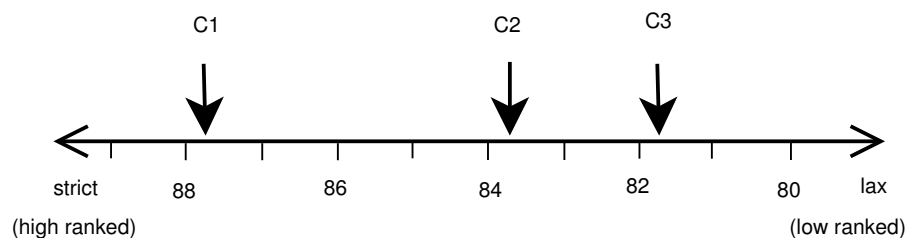
$$(22) \quad (C_2 \& C_3) \gg C_1 \gg C_2 \gg C_3$$

1.3.3 Stochastic OT

Classic OT assumes a strict ranking of constraints (21). Given this strict ranking, C_3 can never dominate C_1 or C_2 . Candidates which violate constraints in strata other than the optimal candidate will never surface. Variation is thus only possible if the variants have the same violation profile, i.e. only vary within a stratum, as in table 1.5. In this case we find free variation. However, we will see that other distributions are observed besides the fifty-fifty of free variation. These can be modeled with a stochastic implementation of OT (Boersma and Hayes, 2001).

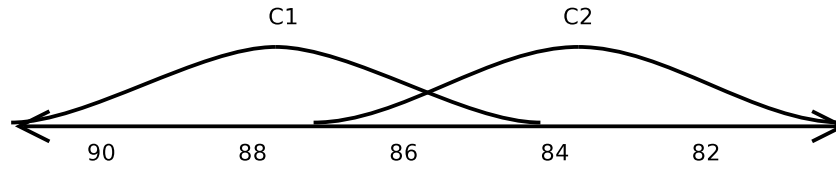
Boersma and Hayes interpret constraint ranking in a different way. First, a linear scale is adopted, as shown in (23). The higher the numerical value of a constraint, the higher the ranking. This allows one to express distances between constraints, e.g. C_1 outranks C_2 more than C_2 outranks C_3 .

(23) Categorical ranking on a continuous scale



Furthermore, the candidates are evaluated stochastically. Whenever a candidate set is evaluated, the exact position of a constraint on the scale is determined. This exact numeric value depends on its ranking, but is perturbed by a random variable. This perturbation is a model of noise in the system. The range of possible selection points for a certain constraint is interpreted as a normal probability distribution with the peak at its ranking value. This is illustrated in (24). A small area in the diagram is enclosed by both curves. In this area, it is possible for C_2 to outrank C_1 . These alternative rankings may give rise to alternative optimal candidates. The probability of the alternative ranking is thus an inverse function of the distance between two constraints. This probability quickly approaches zero if constraints are further away from each other, modeling categorical distinctions.

(24) Stochastic evaluation of constraint ranking



1.3.4 OT and LFG

Although OT has its roots in phonology, it has been applied to other fields of linguistics, including syntax. The application of OT to the field of syntax raises an important question: what is the input? In phonology, the input was assumed to be the underlying form of a word (for production) or its surface form (for interpretation). But what is the underlying form in syntax? LFG-OT has a straightforward answer to this question: the underlying form is an underspecified f-structure. This f-structure represents the main grammatical information that a given utterance expresses, but abstracts away from morphological or lexical (language specific) information.

GEN is then a function from underspecified f-structures to pairs of c-structures and the corresponding (fully specified) f-structures. The candidates' f-structures are all subsumed by the input f-structure. Faithfulness constraints relate the input f-structure to the output f-structure, markedness constraints and alignment constraint determine the shape of the c-structure.

For illustration, we included a (slightly edited) example from Kuhn (2002) in fig. 1.6. It shows how the input f-structure for a simple question is mapped to an infinite number of candidates, of which only three are depicted in the illustration. The candidates are pairs of c-structures and f-structures. Each pair violates some of the relevant constraints in (25)-(27), which were coined in Bresnan (2000).

- (25) OP-SPEC: An operator must be the value of a DF [discourse function] in the f-structure.
- (26) OB-HD: Every projected category has a lexically filled [extended] head.
- (27) STAY: Categories dominate their extended heads.

The first candidate violates OP-SPEC, as the question operator is not associated with any discourse function. *Will* functions as the extended head of *I'* in candidate 3, but cannot function as the extended head of *C'* in candidate 2, as all nodes dominating the extended head should also dominate the projection. Thus Candidate 2, but not candidate 3, violates OB-HD. Candidate 3 does violate the constraint STAY, as *I'* does not dominate its (extended)

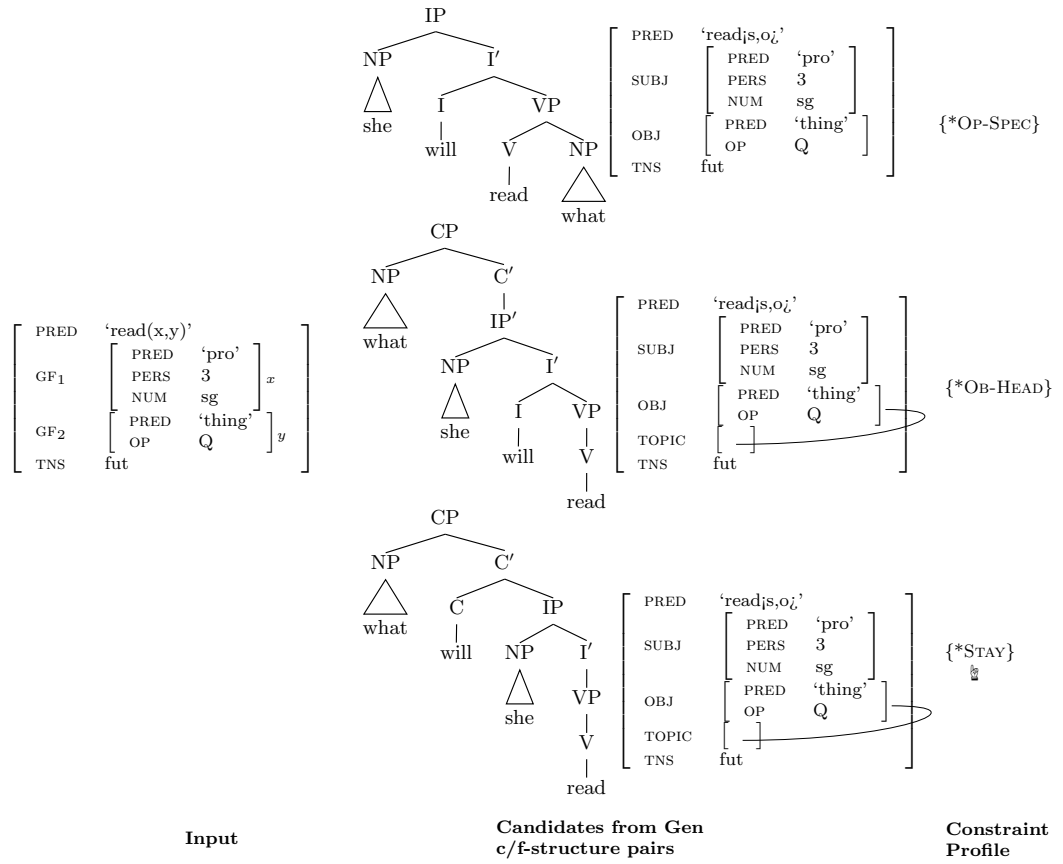


Figure 1.6: General architecture of the OT-LFG framework.

head. Given that the ranking of these universal constraints in English is OP-SPEC≫OB-HD≫STAY, the third candidate is the optimal one.

The OT-LFG framework is relatively new, and various aspects of the framework are still under discussion, including the role of the lexicon and the function of GEN. Bresnan advocates a version of OT-LFG in which the candidates do not contain lexical information Bresnan (2000, 1999, 2002, 2001a). Instead, they are bundles of features, which are decorated with lexical material only after optimization. This ensures that the principle of Richness of the Base is left intact. Van der Beek and Bouma (2004) claim that such an interpretation fails to account for various linguistic phenomena, and instead argue for a function GEN which maps an input *and a language particular lexicon* to a set of output candidates.

In classic OT, the only source of cross-linguistic variation is the ranking of the constraints. Depending on the view on the lexicon, this may or may not be augmented with lexical differences. In both cases, the question arises where that leaves GEN. Radically different answers have been formulated to this question. Kuhn (2003) describes an implementation of an OT-LFG system. The base grammar GEN includes all and only universally inviolable constraints on syntactic structure. In practice, these are the principles of X-bar theory. This leaves intact the principle of a universal GEN (although a language particular base lexicon is used, in line with van der Beek and Bouma (2004), and the candidates are thus language particular). Frank et al. (2001) describe a radically different, but OT-LFG based syntax model. They allow for a full-fledged language particular LFG grammar. This grammar is complemented with a system for marking a candidate parse as good or bad, based on a ranked set of constraints. Evaluation then proceeds as usual: the candidate with the least important constraint violations wins and is considered grammatical. For an elaborate discussion of the differences and similarities between this variant and ‘classic’ OT-LFG, we refer to the original paper (Frank et al., 2001).

We have seen various different ways of modeling violable syntactic constraints. In OT and OT-LFG it is a core feature of the framework. In other systems, such as the Alpino parser, violability is restricted to the probabilistic disambiguation module. These are merely different ways of modeling the same concept: not all constraints in natural language are hard constraints. Johnson (2002) for example shows that categorical OT-LFG systems are very similar to probabilistic language models such as Maximum Entropy Models. In this thesis, we model violable constraints as OT constraints. Such a model allows us to illustrate and control the interaction of various constraints and it is linguistically more insightful than a probabilistic black box. However,

the same constraints may well be formulated or implemented in stochastic grammar components.

1.4 Overview

Chapter 2 argues that the Dutch it-cleft construction is in fact *two* distinct constructions. Analyses are developed for both types of it-clefts, accounting for the complex agreement facts without violating the strict subject-verb agreement rules of Dutch. Corpus data is used for examples that illustrate the characteristics of the construction, and for counterexamples to alternative analyses. For example, when we argue that the focused element should be analyzed functionally as part of the relative clause in intransitive clefts, we illustrate this with a corpus example of a cleft in which the focus fulfills an argument function of the embedded verb. Examples of clefts with demonstrative subjects contradict the claim that clefts have expletive subjects.

Chapter 3 investigates the different realizations of the Dutch dative alternation. While in English the syntactic category of the recipient and the order of the two objects alternate together (direct object followed by PP recipient or indirect object followed by direct object), the two may alternate separately in Dutch. We investigate the hypothesis that general alignment principles influence the argument order alternation, and the verb lexeme influences the NP/PP alternation. This hypothesis is proved partly wrong: general alignment principles are shown to influence the order of the arguments and the verb lexeme only influences the NP/PP alternation, in line with our hypothesis. But pronominality, often assumed to influence the order of the arguments, also influences the category of the recipient, and weight does not influence the order of the arguments in the midfield. The influence of linguistic factors on the distribution of the alternants in the dative alternation is quantified by means of simple statistics.

Chapter 4 focuses on a phenomenon that has received little attention in the linguistic literature so far: determinerless PPs. It is shown that nouns which usually require a specifier may occur without one in certain PPs. Various types of determinerless PPs are distinguished and possible accounts are sketched. For all these accounts it is necessary to know which nouns and which prepositions may occur in which type of detless PP and what their modification potential is. We automatically extract this information from automatically annotated corpora.

Chapter 5 is concerned with the automatic classification of nouns as countable and/or uncountable. We first develop a corpus-based approach, applying both monolingual and crosslingual classification. We then experi-

ment with a second classification strategy, which is based on EuroWordNet, a semantic ontology. We conclude that the corpus-based method outperforms the ontology based method.

Finally, in chapter 6 we summarize and conclude. We briefly point to some directions for future research.

Chapter 2

Clefts

This chapter is concerned with the *it*-cleft construction in Dutch. We claim that we have to distinguish two types of cleft sentences: the transitive cleft, which focuses a nominal phrase and contains a relative clause, and the intransitive cleft, which may focus a range of different phrases and contains a final complementizer clause. Formal analyses of both types are presented. An earlier version of this chapter appeared as Van der Beek (2003).

2.1 Introduction

Dutch *it*-clefts are a puzzling construction. They consist of the same basic elements as English clefts—the pronoun *het* ‘it’, the verb *zijn* ‘to be’, the focused phrase (c-focus) and a final clause—but agreement is different: if the c-focus is plural, then the copula is plural too, even though the subject is *het* ‘it’ (1-a)¹. This appears to be in conflict with the otherwise strict subject verb agreement in Dutch.

Accounting for the agreement in Dutch clefts is further complicated by the fact that the argument structure of clefts depends on whether or not the c-focus is a pronoun: if the c-focus is a full noun phrase, *het* is in the canonical preverbal subject position and the c-focus in the canonical postverbal object position (1-a)-(1-b). But if the c-focus is pronominal, then it is in subject position and *het* is in object position (1-c), generally.

- (1) a. Het zijn niet de vliegtuigen die mij uit de slaap houden.
it are not the airplanes that me out the sleep keep

¹Unless explicitly stated otherwise, all (grammatical) examples are from corpus data. We used several newspaper corpora, *Volkskrant*, *NRC Handelsblad* and *Algemeen Dagblad*, and the Corpus of Spoken Dutch (CGN) to retrieve the examples. For rare constructions, we used the web as an additional source of data.

- It's not the airplanes that keep me awake.*
- b. Het is immers niet de trainer die kansen voor open
 it is after all not the trainer who chances in front of open
 doel verknalt.
 goal loses
After all, it's not the coach who misses the easy shots.
- c. Ik ben het die dom doet.
 I am it that stupid does
I am the one acting stupid.
- d. Het was op zijn aandringen, dat ik de redactie van de
 it was on his insistence that I the wording of the
 adviesaanvraag [...] zo heb veranderd.
 advice appeal have thus changed
*It was on his insistence that I changed the wording of the appeal
 for advice [...].*
- e. Het is omdat ik dit voorheb, dat ik hem begrijp.
 it is because I this have-before, that I him understand
It is because I have this advantage, that I understand him.

A second challenge for the analysis of Dutch *it*-clefts is the difference between clefts with final relative clauses (1-a)-(1-c) on the one hand and *it*-clefts with final complementizer clauses (1-d)-(1-e) on the other hand. While the final clause in the first three examples is introduced by the plural/common relativizer *die* (1-a)-(1-c), the clause is headed by the complementizer *dat* in the examples (1-d)-(1-e).² A similar contrast can be observed in English, where the *that*-clause can often be analyzed as a relative clause (in which *that* may be replaced by *who* or *which*), but not always. See also Quirk et al. (1985) for the characteristics of the final clause in English cleft sentences.

All examples in (1) contain a subject *het*, a form of *zijn* 'to be', a focused phrase and a clause, and in each sentence, given information is extraposed in order to focus a certain constituent. Despite these similarities, this chapter argues that the Dutch *it*-cleft is in fact two constructions: one with the transitive (specificational) copula and a discontinuous TOPIC for clefting nominals and one with the intransitive (existential) verb *zijn* 'to be' for clefting other syntactic categories (as well as some nominals). The first has a final relative clause, but cannot be reduced to any other relative clause construction. The second has a complementizer clause, but cannot be reduced to any other

²The complementizer is homonymous to the neuter relativizer. We know that it is in fact the complementizer because there are no neuter nominal 'traces' in the embedded clause.

complementizer clause construction. We account for the agreement patterns in (1) without violating the generally assumed canonical word order rules for Dutch nor subject-verb agreement. In addition, we show how both the argument structures in (1-b) and in (1-c) can be generated by one set of rules.

Section 2.2 presents an analysis of transitive clefts and in section 2.3 it is argued that the intransitive cleft is a separate construction, for which a separate formal analysis is presented. The chapter concludes with a summary and discussion of some open ends in section 2.4.

2.2 Transitive Clefts

The first type of cleft has a final relative clause and a nominal c-focus, which is either a pronoun or a full NP (1-a)-(1-c). The construction has various interesting features: it appears to violate the otherwise strict subject verb agreement, the relative clause appears not to agree with its antecedent if this antecedent is a pronoun and the argument structure depends on the syntactic category of the c-focus.

We shall show that the final clause cannot be reduced to a regular post-nominal or extraposed relative clause modifier. Instead, it must be analyzed as a specific construction for focusing nominals. After investigating the syntactic properties of the c-focus, the subject pronoun and the relative clause, a formal analysis of the construction is presented which accounts for the agreement features without violating the general word order principles of Dutch or the principle of subject-verb agreement.

2.2.1 Differences between cleft clauses and other relative clauses

The relative clause cleft is very similar to predicative copular constructions in which the NP predicate has a postnominal relative clause modifier. This may even lead to ambiguities between the two readings. Compare the two text fragments in (2-a) and (2-b). Both examples contain an almost exact repetition of example (1-b). In (2-a), it is presented in the context of (1-b) in the corpus. This is an example of a cleft construction. It negates the identification of the person who misses the easy shots with the coach, while putting heavy focus on *trainer*. No other part of the sentence is or can be focused. The c-focus and relative clause do not form a syntactic or semantic unit.

- (2) a. [Coach Vonk] weigert de verantwoordelijkheid voor de
 coach Vonk refuses the responsibility for the
 malaise op zich te nemen. Het is immers niet de trainer die
 slump on self to take it is after all not the coach who
 kansen voor open doel verknalt.
 chances in front of open goal misses
*Coach Vonk refuses to take the responsibility for the slump. After
 all, it's not the coach who misses the easy shots.*
- b. Je zal verbaasd zijn te horen wie er ontslagen is. Het is
 you will surprised be to hear who there fired is it is
 niet de trainer die kansen voor open doel verknalt.
 not the coach who chances in front of open goal misses
*You'll be surprised to hear who got fired. It is not the trainer who
 misses the easy shots.*

In the second fragment, the same sentence is placed in a different context. The prosody of the sentence changes: the main stress shifts from *trainer* to *niet*. Furthermore, the meaning is completely different: the (negated) identification is not between the person who misses the easy shots and the coach, but between some third person (the one who got fired) and the coach, of which we know he missed some easy shots. And in contrast to fragment (2-a), the information structure in this sentence is not fixed: though less likely, the focus (and therefore the main stress) could also be on *open*, *goal* or *misses*. Finally, the two differ with respect to their syntax: the relative clause and the predicate nominal in this sentence form a semantic and a syntactic unit (an NP). This second fragment is not a cleft sentence.

The string may contain some clues as to whether an example sentence is a regular post-nominal modifier or a cleft. For example, proper names and pronouns seldom have relative clause modifiers, but they do occur frequently in clefts. Disambiguation is also possible on the basis of constituent structure: if the complement and the relative clause may be topicalized together, as in (3), without a change of meaning, then the two form one NP and the original sentence is not a cleft construction. But sometimes disambiguation is just a matter of interpretation. The corpus examples in this chapter were only included if their contexts showed them to be clear examples of the *it*-cleft construction, not post-nominal modifiers.

- (3) De trainer die kansen voor open doel verknalt is het niet
 the coach who misses in front of open goal misses is it not
It's not the trainer who misses the easy shots.

For differences between the *it*-cleft and other constructions containing an integrated relative clause in English, see Huddleston and Pullum (2002, p.1416). There is an extensive literature on the semantic and pragmatic characteristics that are specific to the cleft construction. Some pointers are Chomsky (1972); Prince (1978); Delahunty (1981); Atlas and Levinson (1981); Declerck (1988).

There is another construction with a relative clause which is superficially similar to the transitive cleft: the extraposed relative clause modifier. Like in cleft sentences, the phrase immediately preceding the relative clause is not its antecedent (4).

- (4) De gemeente wil namelijk een breed pad aanleggen dat
 the municipality wants namely a wide path build that
 verbonden wordt met de openbare weg.
 connected becomes with the public road
Because the municipality wants to build a wide path that will be connected with the public road.

However, the transitive cleft and the regular extraposed relative clause differ on various points. First of all, while relative clause extraposition is never obligatory, the cleft clause is always extraposed: no non-extraposed variant of the *it*-cleft exists (5).

- (5) *Het die kansen voor open doel mist is immers niet de
 it that chances in front of open goal misses is after all not the
 trainer.
 coach

Furthermore, relative clause extraposition is not restricted to the pronouns *het*, *dit* and *dat*, but clefts are. Similarly, the cleft constructions is restricted to copulas, whereas relative clause extraposition is freely occurs with any verb.

We conclude that *it*-clefts with relative clauses are a construction distinct from other relative clause constructions, which calls for an analysis. In this section, we discuss the syntactic features of the different components of the construction and some previous analyses and finally present a new analysis for this construction which accounts for its characteristics and in particular the Dutch agreement facts while respecting the main principles of Dutch grammar, such as the canonical word order rules and subject-verb agreement.

2.2.2 The c-focus

The two main characteristics of the c-focus are that it is an NP and not (necessarily) predicative: proper names and pronouns can and do appear in this position. If the c-focus is an NP, it takes the complement function. We do not discuss the role of the complement of the transitive copula in this thesis. We will call it OBJ, even though we realize that it is not a regular object, e.g. it cannot passivize.

The fact that the relative clause cleft is restricted to NPs only should not be taken to mean that *all* nominals appear in clefts with relative clauses. Although the large majority of NPs combines exclusively with a relative clause, predicative nominals are not allowed in this construction. More generally, it appears that bare singular nouns combine with *dat*-clauses (6-a) instead of relative clauses (6-b). Interestingly, these bare nominals also allow for free relative extraposition (6-c), which is otherwise not possible, generally (6-f). Once combined with an article, the nominals behave like regular NPs and form relative clause clefts (6-d), but not complementizer clause clefts (6-e). Further research should be carried out to determine exactly which semantic feature is responsible for this exceptional behaviour.³

- (6) a. Het is vooral olie dat ze uitvoeren.
it is mainly oil COMP they export
It's mainly oil that they export.
- b. *Het is vooral olie die ze uitvoeren.
it was mainly oil REL they export
- c. Het is vooral olie wat ze uitvoeren.
it is mainly oil FREL they export
It's mainly oil what they export.
- d. Het was vooral de olie die ervoor zorgde dat de weg
it was mainly the oil REL for it caused that the road
gevaarlijk werd.
dangerous became
It was mainly the oil that caused the road to be dangerous.
- e. *Het was vooral de olie dat ervoor zorgde dat de weg
it was mainly the oil COMP for it caused that the road
gevaarlijk werd.
dangerous became

³It appears that some speakers also allow non-subject forms of pronouns in complementizer clefts with non-subject “gaps” in the clause. We leave these examples aside here.

- f. ??Het was vooral de olie wat ervoor zorgde dat de weg
 it was mainly the oil FREL for it caused that the road
 gevaarlijk werd.
 dangerous became

As we saw in example (1-c), the argument structure shifts if the c-focus is a pronoun. In this case, the c-focus functions as the subject of the cleft sentence: it is in subject position and it is in the subject form. However, there are some interesting exceptions to the rule that pronouns are in subject position. Those pronouns that agree with *is* or *zijn* are occasionally found post-verbally (7-a). This is strictly ungrammatical with first and second person singular pronouns, which take the verb forms *ben* and *bent* (7-b). We do not account for this pattern here.

- (7) a. Maar het zijn wij die iets van jullie kunnen leren.
 but it are we that something from you can learn
But it's us who can learn something from you.
 b. *Maar het ben/is ik die iets van jullie kan leren.
 but it am/is I that something from you can learn

2.2.3 Agreement

In Dutch, the verb agrees with the subject in number and person. Example (1-c) shows that this is also the case in clefts: the nominative first person singular pronoun is in the sentence initial subject position and the verb shows first person singular agreement.

If we replace the pronominal c-focus in (1-c) with a full NP, the argument structure changes. The c-focus is now in object position and *het* 'it' is in subject position (1-b). Now that *het* is the subject, we expect the copula to show third person singular agreement, but surprisingly, this is not the case: if the c-focus NP is plural, the copula is plural too (1-a).

In order to account for these agreements facts, let's first look at the syntactic properties of the pronoun. The pronoun *het* in cleft constructions has often been analyzed as the expletive pronoun (Smits, 1989, for example). If this is correct, then we expect that it is impossible to replace *het* with a demonstrative, which cannot be expletive. However, we do find examples of cleft sentences with demonstratives. The examples are infrequent and mainly found in spoken Dutch, but nevertheless grammatical (8).⁴ This is similar to the German cleft construction, which also allows for a demonstrative pronoun

⁴In the following, whatever we say about the agreement features of *het* 'it' also applies to the pronouns *dat* 'that' and *dit* 'this'.

instead of the German pronoun *es* ‘it’ (Smits, 1989). Also for English, it has been claimed that so-called *th*-clefts are not impossible (Hedberg, 2000). We thus assume that *het* is not an expletive subject. The standard tests for expletiveness, such as the possibility to stress the word or to have emphatic reflexives, fail to distinguish between referential and expletive uses of *het*, as it is a weak and obligatorily stressless pronoun, but do not falsify our assumption that the pronoun is not expletive.

- (8) Goh dat is mezelf die ik hoor
 gosh that is myself that I hear
Gosh, it's me that I hear

Hedberg (2000) also advocates a non-expletive *it* in *it*-clefts. She argues that the pronoun (together with the final clause) functions as a definite description. Furthermore, Gundel (1976) argues for a non-expletive subject in cleft constructions based on data from Russian. Russian does not have expletive subjects, but it does have *eto* ‘it’ in cleft sentences.⁵ Finally, we will see below that there are remarkable similarities between the use of *het*, *dit* and *dat* in cleft sentences and their use in referential simplex copular sentences (also known as truncated clefts). Our hypothesis that the subject pronoun in NP *it*-clefts is different from the commonly assumed expletive pronoun is thus supported by previous work on cleft constructions in other languages, as well as by crosslingual and in-language data.

Secondly, we have to determine the syntactic role of the pronoun. Is *het* really the subject? Dutch has a clear canonical word order, and the pronoun is in the canonical subject position, but various arguments and adjuncts can appear in the canonical sentence initial subject position by means of topicalization. If *het* in example (1-a) is in fact the topicalized object and the plural NP is the subject, then the plural agreement on the verb would be in accordance with subject verb agreement. This analysis fails for multiple reasons. In the first place, topicalized objects must be stressed and *het* is obligatorily unstressed. Therefore, the object pronoun *het* cannot undergo topicalization. Secondly, embedded clauses do not allow topicalization. If *het* were the object, we would expect it not to show up in the subject position

⁵The relevant data are in (i). Note that the Russian construction does not contain a relativizer. Thanks to Lev Blumenfeld and Dmitry Kochenov for sharing their intuitions with me.

- (i) Eto Ivan mne pozvonil
 it Ivan me called
It was Ivan who called me

immediately following the complementizer in embedded clauses. But it turns out that the pronoun does occur in this position (9), if the c-focus is a NP. *Het* is in object position in clefts with a pronominal c-focus, just like in main clauses (10).

- (9) Ze zijn er inmiddels van doordrongen dat het de ondernemers
 they are there by now of convinced that it the producers
 zijn die de welvaart voor het volk creëren.
 that the prosperity for the people create
By now they are convinced that it is the producers that bring prosperity to the people.
- (10) Hij herkent de man en weet dat hij het is die hem
 he recognizes the man and knows that he it is that him
 binnenkort het land uit wil jagen.
 soon the country out wants chase
He recognizes the man and knows that it's him who wants to chase him out of the country shortly.

Additional evidence for the subject-hood of *het* can be found in raising constructions, where the main verb functionally controls the subject of the embedded verb. If we assume that the pronoun *het* is the subject of the cleft sentence, then it should be possible to raise it if we embed the cleft in a raising construction. And we do indeed find such raised cleft constructions (11-a). Recall that in clefts with a pronominal c-focus, *het* was the object and the c-focus was the subject. Although we did not find any examples in our corpus, it does appear possible to raise the focused pronoun when we embed (1-c) under a raising verb (11-b).

- (11) a. Het lijken vooral dit soort instellingen te zijn die in de
 it appear mostly this type organizations to be that in the
 problemen zijn geraakt.
 problems are come
It appears to be mostly this type of organization that came into trouble.
- b. Dus toen leek ik het te zijn, die stom deed
 thus then appeared I it to be who stupid did
So at that point it appeared to be me who was acting stupid.

Now that we have established that the pronoun *het* is the subject, how can we account for the agreement features of the verb? The examples (9) and (11-a) illustrate that both the embedded copula and the raising verb show plural

agreement. Following the strict subject verb agreement in Dutch, we have to conclude that *het* is plural in the examples (1-a), (9) and (11-a), similar to the analysis of *there* in phrase structure grammars (Pollard and Sag, 1994).

There is independent motivation for the existence of a plural and/or common *het/dat/dit* ‘it/that/this’. The distribution of these pronouns is not restricted to clefts and raising constructions: they also show up in other types of copular sentences, both as personal pronouns (12-a) and resumptive pronouns (12-b) (see also Rullman and Zwart (1996)). A classic discussion in Dutch linguistics deals with the question which of the constituents in sentences like (12-a) is the subject (Merckens, 1961; Bos, 1961), where the word order suggests that *dat* is the subject, but subject verb agreement suggests that *soldaten* ‘soldiers’ is the subject. It is possible to analyze the pronoun as the subject (in accordance with the Dutch word order rules) and account for the plural agreement on the verb if the pronoun has a plural value for NUM.

- (12) a. Dat zijn pas echte soldaten.
 that are now real soldiers
 Now those are real soldiers.
 b. VUT’ers, DOP’ers - dat zijn vroeg gepensioneerden en
 VUT-ers DOP-ers - that are early retired ones and
 economisch zelfstandigen.
 economical independent ones
 VUT-ers, DOP-ers, those are the early retired and the economically independent.

In these examples, the pronoun itself does not show agreement, but subject verb agreement in example (12-a) and resumptive pronoun antecedent agreement in example (12-b) indicate that the value for NUMBER on the subject is in fact plural (and GENDER is common). Based on the fact that some pronouns can have both singular and plural number and both neuter and common gender, one may think that these pronouns are simply underspecified and the finite verb has defining equations specifying its subject’s agreement features (13-a).

- (13) a. *zijn*: V (↑PRED) = ‘be-equal-to((↑SUBJ)(↑OBJ1))’
 (↑SUBJ PERS) = 3
 (↑SUBJ NUM) = pl

This would explain the grammaticality of the examples (12-a) and (12-b) but not the ungrammaticality of (14-a) and (14-b), in which the use of a “non-agreeing” pronoun is not allowed.⁶ Given the entry in (13-a), nothing prevents unification of the underspecified subject pronoun and the defining subject specifications on the verb. The ungrammaticality of these examples follows naturally if we assume instead that the agreement constraints on the verb are constraining equations (13-b): the subject pronoun is not specified for number and person and therefore the constraining equations on the verb are not satisfied. But if the agreement equations are constraining, how do we explain the grammatical and apparently non-agreeing examples? To answer this question, we look at the restrictions on the use of plural/common *het*.

- There are two ways of using the pronouns *het*, *dit* and *dat* grammatically as subjects of a simple copular clause. As we saw earlier, the pronouns may be used as singular or plural subjects of a copular clause with a nominal object (12-a), (12-b). If no nominal object is present, the pronoun *het* ‘it’, the resumptive pronoun *dat* ‘that’ and the demonstratives *dit* ‘this’ and *dat* ‘that’ are still possible, but only if they are singular and neuter (15-a). In other words: the apparent non-agreement is only possible with a nominal object.

- This pattern can be accounted for if we assume that the pronoun has optional default agreement features and agrees in number and gender with the copular

⁶The examples are fine if an explicitly plural pronoun such as *die* ‘those’ is used instead of *dat* ‘that’.

complement: if this complement is a nominal phrase, then it shares its agreement features with the pronoun, possibly causing the pronoun to be plural and/or common. If the complement is adjectival, this sharing is impossible, because the adjective is not defined for number or gender. As a result, the pronoun cannot be made plural or common and only can be used with its “default” values, neuter singular. We assume a lexical entry for *het* as in (16)⁷ and entry (13-b) for *zijn*. The specifications for number and gender are optional, so that they can be overridden by the agreement features of a nominal complement. In case the subject cannot ‘get’ agreement values from the (adjectival) predicate, it can only satisfy the constraining equations on the verb by instantiating the default value for number and person: singular neuter. This explains why the examples (14-a), (14-b) and (15-b) are out, but (12-a), (12-b) and (15-a) are ok.

- (16) *het*: PN (↑PRED) ‘pro’
 (↑PERS) 3
 ((↑NUM) sg)
 ((↑GEN) neut)
 (↑PRONTYPE) cop

In addition, the pronoun has a feature PRONTYPE with value ‘cop’ (copular). This feature-value pair sets *het* ‘it’, *dat* ‘that’ and *dit* ‘this’ apart from all other pronouns. It reflects the fact that these three pronouns form a distinct class with a specific syntactic distribution and semantics (Declerck, 1988).

Subject-complement agreement in number and gender is not observed in all Dutch copular sentences. Number agreement is widespread, but there are exceptions, such as bare singular nouns which are used to predicate over plural subjects and sentences like the following example from the web (17), where the number is not shared.

- (17) dat als zij mij waren, ze SPF niet zouden noemen.
 that if they me were they SPF not would mention
 that they wouldn’t mention SPF if they were me.

Gender agreement across the copula exceptional in Dutch. Nouns have a fixed, lexical specification GEN, which makes it impossible to ‘adjust’ gender in order to agree with the subject. As a consequence, gender mismatches between subject and complement in copular sentences (18) are very common.

⁷The parentheses around the optional features translate to the following disjunction: (↑NUM) ∨ (↑NUM)=sg and (↑GEN) ∨ (↑GEN)=neut.

- (18) Het knelpunt is de ondoorzichtigheid.
 the bottleneck_{neuter} is the opacity_{common}
The bottleneck is the opacity.

Only some pronouns have flexible GEN specifications: demonstratives have two forms, one for neuter and one for common gender (e.g. *dat* ‘that_{neuter}’ vs. *die* ‘that_{common}’). Here we do in fact find the expected contrast in referring expressions, as illustrated in the constructed example (19). But we saw that we get the copular pronouns *het*, *dit* and *dat* (in which the gender contrast is not observable on the surface) in sentences with nominal predicates. In addition, the common gender pronoun *die* is becoming more and more acceptable referring to neuter objects. Something similar is happening with the personal pronoun *hem* ‘him’, which is also used for both neuter and common objects. In short, although gender agreement within the NP (e.g. between the article and the noun, concord agreement) is very strict in Dutch, gender agreement between NPs (index agreement) is rare.

- (19) a. Ik heb mijn oude broek weggedaan. Die was
 I have my old trousers_{sg,comm} thrown-away that_{comm} was
 inmiddels te klein geworden.
 by-now too small become
I have thrown out my old trousers. By now they have become too small.
- b. *Ik heb mijn oude broek weggedaan. Dat was
 I have my old trousers_{sg,comm} thrown-away that_{neut} was
 inmiddels te klein.
 by-now too small become
- c. Ik heb mijn oude overhemd weggedaan. Dat was
 I have my old shirt_{sg,neut} thrown-away that_{neut} was
 inmiddels te klein geworden.
 by-now too small become
I have thrown out my old shirt. By now it has become too small.
- d. ?Ik heb mijn oude overhemd weggedaan. Die was
 I have my old shirt_{sg,neut} thrown-away that_{comm} was
 inmiddels te klein geworden.
 by-now too small become

One may argue that Dutch does have a general principle of subject-predicate agreement in copular sentences, but it only shows in cases where we have a pronoun. This can be modeled with a violable constraint in the Optimality Theoretic (OT) tradition. This constraint on agreement should

be outranked by a faithfulness constraint stating that lexical gender specifications should be faithfully realized: only some pronouns can satisfy both constraints and in all other cases we will get a violation of the lower ranked constraint, resulting in a gender mismatch. Alternatively, one may say that this is a peculiarity of those pronouns. This can be modeled by encoding the agreement constraint on the lexical entries of the pronouns.⁸ Our analysis is compatible with both an OT-style approach and pronoun specific functional annotations.

2.2.4 The relative clause

Clefts with a nominal c-focus have a final relative clause.⁹ The relativizer appears to agree in gender with the c-focus: *die* for common singular nouns and plurals and *dat* for neuter singular. It would nevertheless be incorrect to state that the clefted element is the antecedent, because the embedded verb does not agree in person with the c-focus (1-c), as it does in adjoined relative clauses (20).¹⁰ In section 2.2.1 it was furthermore noted that cleft clauses differ from relative clause modifiers of the predicate with respect to prosody, semantics and pragmatics.

- (20) En ik, die dit vertel ben Tina.
 And I, who this tell_{1sg} am Tina
And I, who tell this, am Tina

Alternatively, one could assume that the object is the antecedent. This gives the same results in most cases, since the object and the FOCUS usually coincide. But not if the c-focus is a pronoun. In that case the object (and

⁸The annotation needed to put this constraint on the copular pronoun is rather complex, as it needs to account for both argument structures: $(\uparrow\text{NUM})=((\text{SUBJ}\uparrow)\text{OBJ NUM}) \vee ((\text{OBJ}\uparrow)\text{SUBJ NUM})$ (and a similar one for gender). This constraint states that the number value of the pronoun is unified with the number value of the object if the pronoun is the subject, and NUM is unified with the number value of the subject if the pronoun is the object.

⁹We assume an analysis of relative clauses along the lines of Dalrymple (2001) and Falk (2001): the relative clause is a headless CP with the relative pronoun in SpecCP. The fronted phrase is the TOPIC of the embedded clause and the f-structure of the relative pronoun is the value of a feature RELPRO

¹⁰In old-Dutch and in some bible texts one can also find first person verbs in *it*-clefts. We do not account for these archaic examples here.

- (i) Ik ben het die uw overtredingen uitdelg om mijnentwil.
 I am it that your transgressions take-away_{1sg} for my-wish
It is me that takes away your transgressions because that is my wish.

thus the antecedent) is *het*. The third person agreement on the embedded verb in (1-c) would thereby be explained.

Example (21) appears to be a counterexample to this analysis: the embedded verb is plural, whereas the antecedent is *het*. Similarly, the relativizer in (22) is of common gender, while the antecedent is *het*. However, with the lexical entry proposed in (16), these examples are no longer problematic, as *het* unifies its number feature with that of the pronoun *we* ‘we’.

- (21) Wij zijn het die alle partijen bij de les moeten houden
 we are it who all parties at the lesson must hold
It's us who should make sure all parties stay focused
- (22) Hij was het ook die P.J.H. Cuypers in de arm nam [...].
 he was it too that P.J.H. Cuypers in the arm took
It was him, too, who got P.J.H. Cuypers involved.

One disadvantage of the object antecedent approach remains: its discourse function. The cleft construction is a focus construction. It focuses one element (the clefted element or c-focus, mapped to FOCUS in f-structure), while the given or backgrounded information is extraposed. Under the object antecedent analysis, the old information from the clause is analyzed as a modifier of the object, which in most cases is the c-focus. Thus, it will be part of FOCUS. This is in contradiction with it being given or background information. Furthermore, the information structure of clefts is assumed to be the same, irrespective of the syntactic category of the c-focus. But under the object antecedent analysis, the clause is part of FOCUS if the c-focus is a full NP but not if c-focus is pronominal (because the c-focus is the subject in that case). This makes the object antecedent approach an unattractive analysis.

The fact that it is difficult to find an antecedent for the relative clause, has led to the hypothesis that there is no antecedent and the relative clause is a free relative. Akmajian (1970) analyzed English clefts as pseudo-clefts that had undergone a transformation, moving the free relative to the right edge. A closely related analysis was presented for Dutch in Van der Beek (2001). There, the extraposed clause is analyzed as a free relative clause that is extraposed by means of independently motivated extraposition rules. The analysis of the final clause as a free relative accounts for the agreement facts: if the free relative is in fact the extraposed subject, then the plural free relative in (1-a) does agree with the plural verb.

An important counterargument to free relative accounts is that the form of the relative clause is not the same as a free relative: instead of the relativizers *die* and *dat* for common and neuter antecedents, free relatives use

wie and *wat* for free relatives referring to animate versus inanimate objects.¹¹ Furthermore, free relatives are always singular, with a universal or exhaustive reading, whereas the final clause of a cleft can be plural (1-a).

A second problem for the free relative analysis is that extraposition of free relatives involves expletive insertion. Both the free relative and the expletive map to the same argument function, so that the requirements of both coherence and completeness are met. As we have seen, there is reason to believe that the pronoun is in fact *not* expletive. This means that *het* has a PRED feature, which would clash with the PRED value of the extraposed subject under the free relative analysis.

The relative clause is not a modifier of the OBJ or FOCUS and it is not a free relative. That leaves two possible analyses: the antecedent of the final clause is either the SUBJ or TOPIC. The subject antecedent analysis was first suggested for English by Jespersen (1927). According to this analysis, the final clause is a relative clause that restricts the interpretation of *it*. In English, this is always both subject and topic. Jespersen developed his analysis for English and thus does not address the Dutch agreement pattern: it does not follow from this analysis that the relativizer obligatorily has the same gender as the clefted element in Dutch nor that the verb in example (1-a) should be plural. With the lexical entry for *het* presented in (16) and SUBJ-OBJ agreement in copular constructions, the agreement pattern could be accounted for. But the Jespersen analysis has the same disadvantage as the object antecedent analysis: the discourse function of the relative clause would vary.

That brings us to the analysis we propose in this chapter: the relative clause as a modifier of the topic pronoun *het*. This analysis predicts the correct NUM and GEN values if we combine it with the lexical entry for *het* discussed before. The NUM and GEN values of the pronoun unify with those of the object. The verb can now check for the appropriate values on the subject, which is either the pronominal c-focus or the topic pronoun *het* with the unified agreement features of the object. The agreement between the relativizer and the antecedent is also unproblematic under this analysis, because the antecedent *het* now has the same agreement features as the c-focus. The relative clause is always a part of TOPIC. This nicely reflects the observation that the information in the final clause of a cleft has to be given (Declerck, 1988). Our analysis resembles the analysis in Hedberg (2000), who claims that the pronoun and the relative clause function as a discontinuous definite

¹¹But note that the reference grammar of Dutch Haeseryn et al. (1997) does allow *die* and *dat* as the heads of free relatives, although the non-cleft examples are marginal. In addition, the dictionary of Dutch from 1500-1976 does list them as possible heads of free relatives (de Vries et al. (1882-1998), column 2517-2518)

description. But in her syntactic analysis of the construction, she analyzes the clause as adjoined to the focused phrase. The LFG framework facilitates an account of *it*-clefts in Dutch in which the pronoun and the clause form a unit both semantically and syntactically: the two components map to same f-structure even if they are discontinuous on the level of c-structure. This analysis is formalized in the next section. The construction is treated as a specific focus construction, distinct from regular relative clause extraposition. The idiosyncratic properties of the construction (see also section 2.2.3) can thus be dealt with.

Note, finally, that Dutch has another construction in which an obligatorily clause final relative clause modifies a pronoun. These are instances of the quantitative use of the R-pronoun *er*, where the quantitative element is left out (23). Like in clefts, the pronoun and the extraposed relative clause do not form a constituent on c-structure.

- (23) Maar er zijn er ook die het met achthonderd dollar in
 but there are R-pron also that it with eight hundred dollar in
 de maand moeten doen.
 the month must do
But there are also people who must do with 800 dollars a month.

2.2.5 Formalization

We have argued that the pronoun *het* ‘it’ has a lexical entry with default agreement values. The transitive Dutch *it*-cleft consists of this pronoun, a second nominal argument (the c-focus) and a relative clause. The antecedent of the relative clause is the topic pronoun *het*.

The different parts of the analysis are combined in the c-structure rules in figure 2.3 on page 49. The rules are for main clause clefts. Although the c-structure rules for subordinate clauses are different, the idea is the same: two nominal arguments and a relative clause on the right edge. It is this relative clause that carries the construction specific f-structure specifications for focus on the clefted element and the pronoun *het* with discourse function TOPIC in either subject or object position, bearing a feature ADJ that is filled by the final clause as a whole.¹² An example c-structure is given in fig. 2.1, the corresponding f-structure in fig. 2.2.

Like in regular relative clauses, the relative pronoun in the final clause can be embedded in a PP (24). These examples are automatically accounted

¹²The concept of a sentence final CP that maps to the ADJ of the non-expletive pronoun *it* is also found in Berman’s analysis of extraposed argument clauses in German (Berman, 2001).

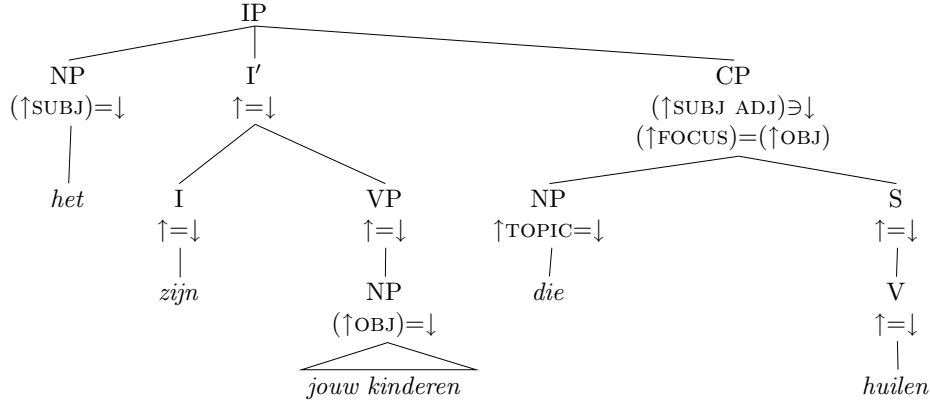


Figure 2.1: C-structure for *Het zijn jouw kinderen die huilen*

for by the regular relative clause rules.

- (24) Het zijn dat soort reacties waardoor de military-ruiters
 it are this kind reactions through-which the military riders
 zich onbegrepen voelen.
 themselves misunderstood feel
It is this type of responses that make the military riders feel misunder-
stood.

The c-structure rules in figure 2.3 show that the transitive *it*-cleft has various construction specific features that have to be stipulated in the c-structure rule: the relative clause does not form a \bar{N} with its antecedent, the relative clause is obligatorily at the right edge and the TOPIC has to be of a particular pronoun type. On the other hand, we used the independently motivated c-structure rules for transitive sentences and specialized them in order also to cover cleft sentences. The only component that was added is the optional relative clause with all the construction specific information.¹³

This analysis leaves the canonical Dutch word order intact: the canonical subject position, filled by *het*, is associated with the grammatical subject function. At the same time, it meets the requirement of subject verb agreement: the pronoun *het* is in fact plural, since it unifies its AGR values with those of the object. This unification also predicts the observed pattern of agreement between the relativizer and the pronoun.

We do not account for the distribution of the two argument structures

¹³The rules for the transitive cleft can be merged with the general rules for transitive clauses by adding the CP optionally to the general IP rule.

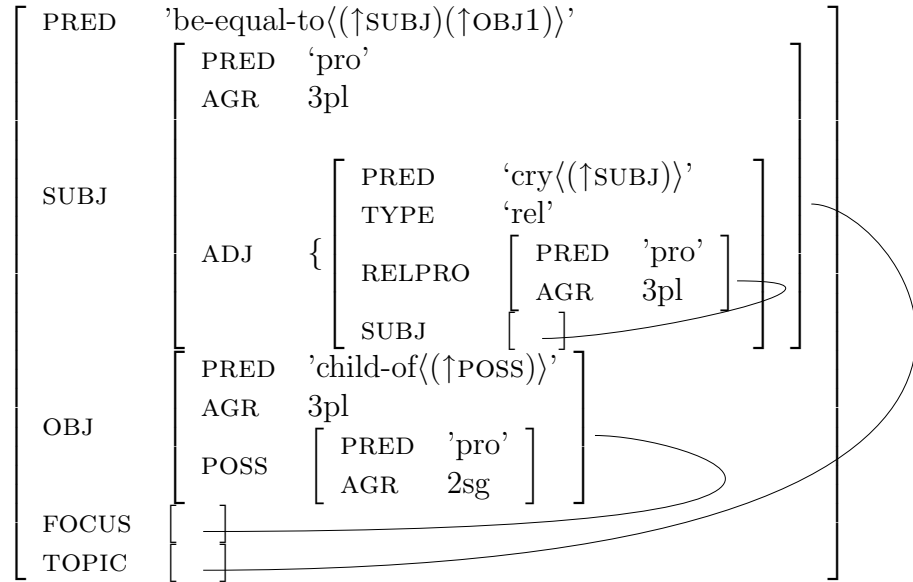
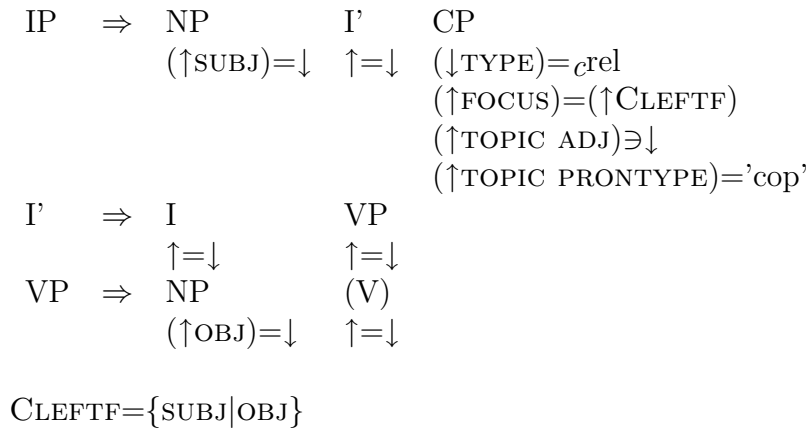
Figure 2.2: F-structure for *Het zijn jouw kinderen die huilen*

Figure 2.3: C-structure rules for nominal clefts

of the transitive cleft in the c-structure rules. The rules in figure 2.3 generate both argument structures for both pronouns and full NPs, even though focused pronouns are virtually always realized as subjects in a cleft construction, and focused NPs are realized as objects in a cleft construction. It is assumed that general constraints penalize copular object pronouns, and focused pronouns in particular. Thus candidates like (25) are excluded.¹⁴

- (25) a. *omdat het HEM is die huilt
 because it him is who cries
 b. *omdat jouw ZOON het is die huilt
 because your son it is who cries

This assumption that the argument structure in clefts is an effect of a more general mechanism is supported by the fact that the same effects can be observed in other copular sentences. Haeseryn et al. (1997) list 3 copular constructions in which pronouns can function as a complement (26). In all cases, the subject is a pronoun, too. Sentences with a full NP subject and a pronominal complement are ungrammatical (27). Apparently, pronouns are realized as subjects whenever possible.¹⁵

Only if both functions are realized by pronouns is a pronominal complement acceptable. The only surprising example is (26-c), where the subject is *het* and the complement a personal pronoun. After all, this argument structure is out in cleft sentences. Coppen (1996) noted that the same string with a neutral stress pattern is ungrammatical (28-a). He accounts for the ungrammaticality of the example based on the assumption that the copular complement is thematically associated with the subject, but receives the non-subject case because of its position. This is no problem for nouns or the pronouns *het*, *dat* and *dit*, which do not carry casemarking, but it is a problem for personal pronouns, the only category in Dutch which does show case marking. He does not account for the contrast between (28-a) and (26-c). For cleft sentences, it is important to note that the stress pattern in (26-c) is not available: stress is on the c-focus.

¹⁴In fact a number of examples like (i) can be found on internet. This shows that the constraint is not categorical. In the next chapter we investigate how one may account for such non-categorical distinctions.

(i) Ik denk dat het hem is die ik bedoelde
 I think that it him is that I meant
 I think it was him that I meant.

¹⁵This assumes that the relation between the two copular arguments in these constructions is symmetrical.

- (26) a. Als ik jou was ...
 if I you were ...
 If I were you ...
 b. Hij is 'm.
 he is him_{reduced}
 He is it.
 c. Het IS 'm
 it IS him_{reduced}
 It IS him.
- (27) *dat kandidaat A jou is.
 that candidate A you is
- (28) a. *Het is hij/hem.
 it is he/him
 b. Hij is het.
 he is it
 It's him.

It appears appropriate to treat the argument structure differences in clefts and other copular constructions as part of a yet more general distributional phenomenon. If we compare the argument functions of *het*, personal pronouns and full NPs, we find striking differences. For all categories, the subject function is most frequent, but this preference is much stronger for pronouns than for NPs. Looking at subjects, direct objects and predicative complements only, we find for NPs the following distribution in the Alpino Treebank: 65% subject, 29% direct object and 6% predicative complement. For *het*, the distribution is 77% subjects vs. 22% direct objects and 1% predicative complements.¹⁶ Finally, of the 2075 relevant personal pronouns, 94% had the subject function and 6% had the direct object function; only one personal pronoun functioned (grammatically) as a predicative complement (29). Although it is technically possible to constrain subject foci to personal pronouns with functional annotations, we assume that our argument order variation is a direct consequence of these more general distributional phenomena.

- (29) Dat was 'm dan, de Puskas van het Poolse voetbal.
 that was him then the Puskas of the Polish soccer
 So that is him, the Puskas of Polish soccer.

¹⁶Discarding sentences with expletive, preliminary subjects and extraposed sentential subjects.

2.3 Intransitive Clefts

So far, we only looked at clefts with a final relative clause. But the clefts in (1-d) and (1-e) do not contain a relative clause. In the next section we discuss this and many other differences between the clefts in (1-a)-(1-c) (transitive clefts) on the one hand, and (1-d)-(1-e) (intransitive clefts) on the other hand.

2.3.1 Differences between transitive and intransitive clefts

The clause The first difference between both types of clefts is the final clause. While transitive clefts have final relative clauses, this second type of cleft has a subordinate final clause headed by the complementizer *dat*. Although this complementizer is homonymous to the neuter singular relativizer, we know that it is in fact a complementizer because there are no neuter singular “traces” in the clause.

The claim that intransitive clefts have a final complementizer clause is not universally agreed upon. Smits (1989) argues that the word *dat* introducing the clause is of a special syntactic category called *relative particle*, which introduces a relative clause. This particle is not only used in intransitive clefts, Smits claims, but also in a specific type of relative clause, namely one that modifies a temporal expression (30).

- (30) Hamills rol [...] in Jay and Silent Bob Strike Back is de eerste
 Hamil’s role in Jay and Silent Bob Strike Back is the first
 keer dat hij de draak steekt met Star Wars.
 time that he makes fun of Star Wars
*Hamil’s role [...] in Jay and Silent Bob Strike Back is the first time
 that he makes fun of Star Wars.*

This is in line with Smits’ (informal) definition of relative clauses as “any construction part of which is a subclause that modifies an expression external to that subclause [...]”. It furthermore allows for the generalization that any cleft has a relative clause. But there are a number of problems with this analysis. First of all, the assumption of a relative particle raises the question why the usage of this particle in clefts is so much broader than in regular relative clauses. Secondly, it is unclear what the antecedent is of the particle in cleft sentences, especially since the clause does not seem to modify any element in the c-focus. Furthermore, one wonders why this relative particle cannot function as an argument in the embedded clause (31), just like relative pronouns. And finally, we will see later on that we do find intransitive cleft

sentences (but not modifiers of temporal expressions) with argument ‘traces’ in the embedded clause. It is unclear how Smits (1989) would account for this contrast.

- (31) *De eerste keer dat ik mij goed herinner was op 5 februari.
 the first time that I REFL well remember was on 5 February

We assume that a word which looks and behaves like a complementizer is in fact a complementizer and introduces a complementizer clause. We thus have to distinguish between clefts with relative clauses and clefts with complementizer clauses.

C-focus Secondly, there are differences with respect to the categories that may be focused. Transitive *it*-clefts only focus nominals, but complementizer clefts may focus a wide range of categories, including PPs (1-d), CPs (1-e) and AdvPs (32-a). *it*-clefts with APs (32-b) have been reported grammatical in the literature Smits (1989), but intuitions differ from one speaker to another and no corpus examples were found. In section 2.2.2, we furthermore saw that a restricted set of nominals occurs not in the relative clause construction, but in the complementizer construction.

- (32) a. Het is daar dat de verveling toeslaat.
 it is there that the boredom attacks
 It's there that boredom attacks.
 b. ?Het is rood, dat hij zijn kamer verft
 it is red that he his room paints
 It's red that he paints his room

The pronoun Another difference between the two constructions is the fact that the pronoun *het* cannot be replaced by a demonstrative (33-a), as it could in transitive clefts. This suggests it is an expletive, as members of like categories are otherwise generally interchangeable.

- (33) a. *Dat is daar dat de verveling toeslaat.
 that is there that the boredom attacks
 b. *HET is daar dat de verveling toeslaat.
 IT is there that boredom attacks
 c. *Hetzelf is daar dat de verveling toeslaat.
 itself is there that the boredom attacks

In section 2.2 it was already shown that other tests for expletiveness fail to make a clear distinction between the expletive *het* and the non-expletive but

obligatorily unstressed pronoun *het*, but there is some further evidence that the pronouns in the two constructions are dissimilar. Recall that Gundel (1976) argued for a non-expletive cleft subject based on data from Russian, which does not have expletive subjects, but does have *eto* ‘it’ in cleft sentences. This appears to contradict our assumption that *het* is an expletive in Dutch complementizer clefts, but the Russian examples all have a nominal focus. It turns out that the *eto*-cleft is in fact only possible with a nominal phrase in focus: different constructions are used to focus PPs or complementizer phrases. Hedberg (2000), which also argued against expletive cleft subjects, explicitly restricts the argument to NP *it*-clefts only. As the large majority of Dutch NP clefts are of the transitive type, this supports our hypothesis of a non-expletive subjects in transitive clefts, without contradicting the expletive status of the subjects in intransitive clefts.

The copula The two types of clefts also differ with respect to the verb. Relative clause clefts may use another copula instead of *zijn* (34-a), while this appears not to be possible for clefts with complementizer clauses (34-b).¹⁷

- (34) a. Toch bleken het uitgerekend de Democraten die [...] het
 Yet appear it calculated the Democrats that [...] the
 meest op hadden met de watersnood.
 most on had with the flooding
*And yet, of all parties, it turned out to be the Democrats who
 cared most about the flooding.*
- b. *Toch bleek het op Democratisch initiatief, dat er
 Yet appeared it on Democratic initiative that there
 steun voor de watersnood kwam.
 support for the flooding came

Argument structure The transitive cleft has two arguments, a subject and an object. This object is usually the c-focus. It is difficult to analyze the complementizer clefts in the same way. The c-focus of an intransitive cleft would make a very unusual object, since it is almost never nominal. And semantically, the construction does not resemble a transitive sentence either: in contrast to the relative clause clefts, the complementizer clefts (35-a) cannot be paraphrased as canonical specificational sentences (35-b), even if we transformed the *that*-clause into a locational free relative (35-c). The best paraphrase would be the simplex sentence in (35-d).

¹⁷Thanks to Frank Van Eynde for pointing this out to me.

- (35) a. Het was in Polen dat het eerste vrije vakverbond onder het
 it was in Poland that the first free union under the
 communisme werd opgericht.
 communism was founded
*It was in Poland that the first free union under communism was
 founded.*
- b. *Dat het eerste vrije vakverbond onder het communisme werd
 that the first free union under the communism was
 opgericht was in Polen.
 founded was in Poland
- c. *Waar het eerste vrije vakverbond onder het communisme
 Where the first free union under the communism
 werd opgericht was in Polen.
 was founded was in Poland
- d. In Polen werd het eerste vrije vakverbond onder het
 in Poland was the first free union under the
 communisme opgericht.
 communism founded
The first free union under communism was founded in Poland.

Finally, the transitive cleft has a variable argument structure and both *is* ‘is’ and *zijn* ‘are’ occur in the matrix clause. In the intransitive cleft, the pronoun is always in subject position and the verb is always singular.

We conclude that the construction that is generally called the *it*-cleft construction consists of two distinct constructions, at least in Dutch: one with a relative clause, and one with a complementizer clause. A similar claim has been made for English clefts by Pinkham and Hankamer (1975). They argue that in English, non-nominal clefts must be derived from simplex sentences such as (35-d) by means of extraction of the focus, creation of the matrix copular construction, and extraposition of the original sentence. Nominal clefts on the other hand are derivationally ambiguous: they may be derived from a simplex sentence in a similar way as non-nominal clefts, or alternatively the copular construction may be base generated with a headless clausal subject, which is then relativized and extraposed, leaving behind an (expletive) pronoun *it*. The two proposals share certain features, e.g. only nominals combine with relative clauses in clefts. But the data, arguments and analyses differ greatly. Pinkham and Hankamer (1975) do not say much about agreement, as agreement in English clefts is always third person singular. The analysis focuses primarily on the status of the verb *to be*, which

is base generated in one derivation but not in the other. From this, they predict facts about connectivity, reflexivization and negation, which do not translate to the Dutch construction. The dual derivation analysis does not look at the differences with respect to the pronoun, the verbs that can be used, or the semantics. Furthermore, the analysis presented in this thesis does not assume the nominal cleft to be derivationally ambiguous, while this is an important feature of the analysis in Pinkham and Hankamer (1975).

Before turning to the analysis of this second type of cleft, we first discuss the differences between complementizer clefts and other complementizer clause constructions.

2.3.2 Differences between intransitive clefts and other complementizer constructions

We have argued that the sentences which are normally considered to be of the same kind (i.e. *it*-clefts) are in fact two distinct constructions and we have proposed an analysis for the first: the typical cleft construction with a nominal c-focus was analyzed as an instance of the transitive (specificational) copula. The question arises what sort of construction the other sentences are. They consist of an expletive subject, the verb *to be* and an extraposed complementizer clause. Two types of sentences contain these same elements: sentences with *that*-clause modifiers and sentences with extraposed clausal subjects, and one may ask if our ‘clefts’ may be grouped under one of them.

That-clause modifiers are clauses headed by the complementizer *dat* that modify a preceding constituent, usually a nominal temporal expression (36-a). In section 2.3.1 we already saw an example of this construction (30). Often, the NP and *that*-clause co-occur with a subject *het* and a copula, as in example (36) below.

- (36) a. Het is de eerste keer dat een minister op deze manier
 it is the first time that a minister on this way
 ingrijpt.
 intervenes
 It is the first time a minister intervenes in this way.
- b. Het is de eerste keer dat een minister op DEZE manier
 it is the first time that a minister on THIS way
 ingrijpt, maar hij heeft vaak genoeg op andere wijze
 intervenes but he has often enough on different manner
 zijn macht doen gelden.
 his power let count

It is the first time the minister intervenes in THIS way, but he has exercised his power in other ways often enough.

Although the resulting structure is similar to our clefts, there are important differences. The pronominal subject in sentences with *that*-clause modifiers is referential: it refers to a situation or an event. As a consequence, it can be replaced by a full NP, as we already saw in section 2.3.1 (30). This is impossible in intransitive clefts.

Secondly, the *that*-clause modifies the preceding phrase and forms a constituent with it. Thus, example (37-a) is a grammatical NP, but (37-b) is not a constituent nor is the nominal example (37-c).

- (37) a. de eerste keer dat een minister op deze manier ingrijpt
 the first time that a minister on this way intervenes
 the first time a minister intervenes in this way
 b. *in Polen dat het eerste vrije vakverbond werd opgericht
 in Poland that the first free union was founded
 c. *olie dat ze uitvoeren
 oil that they export

Furthermore, the focus of the sentence may be inside the modifying *that*-clause (36-b). In relativizer clefts, the subordinate clause is always given and never focused: the focus is on the phrase in the object position.

Another construction that is similar to our cleft examples is clausal subject extraposition. Certain predicative adjectives and nouns allow extraposition of the subject *that*-clause (38).

- (38) a. Het is duidelijk dat het met Abu Ammar is gedaan.
 it is clear that it with Abu Ammar is over
 It is clear that Abu Ammar is over.
 b. Het is van wezenlijk belang dat we nu ingrijpen.
 it is of real importance that we now intervene
 It is of great importance that we intervene now.

These extrapositions are analyzed as transitive copular constructions. The subject is a preliminary, expletive subject and the clause is the ‘real’ subject. Both map to the SUBJ function. Like complementizer clefts, the construction consist of an expletive, a non-verbal complement (usually an adjective or a PP) and a *that*-clause. We nevertheless argue that the two structures are very different from one another and should not be analyzed in the same way.

Our first argument is that the predicate (the adjective or PP) in these extraposition constructions predicates a property of the whole proposition,

whereas the focused phrase in an intransitive cleft specifies a focus within the extraposed clause. Compare on the one hand the intransitive cleft (35-a) and its paraphrase with the focus *in situ* as a modifier of the event (35-d), and on the other hand the propositional predication in (39-a) and the constructed *in situ* variant (39-b), which means something completely different. This difference is reflected by the fact that we do find adverbs in focus in intransitive clefts (32-a), but not in *that*-clause extraction. Similarly, propositional predications may be non-extraposed and without *het* (39-c), but this is not possible for clefts (35-a)-(35-b).

- (39) a. Het is goed dat we de grenzen vaststellen.
 it is good that we the limits determine
 It is good that we determine the limits
 b. We stellen de grenzen goed vast
 we determine the limits good
 We determine the limits well.
 c. Dat we de grenzen vaststellen is goed
 that we the limits determine is good
 It is good that we determine the limits

Secondly, *it*-clefts only occur with *zijn*, but we find occurrences of *that*-clause extrapositions with verbs other than *zijn* ‘to be’ (40)-(41).

- (40) Maar het lijkt evident, dat Endel hem ook hard nodig heeft.
 but it seems evident that Endel him also hard need have
 But it seems clear that Endel seriously needs him as well.
 (41) Ik vind het verstandig dat Inge dit toernooi laat schieten.
 I find it wise that Inge this tournament let shoot
 I think it's wise that Inge skips this tournament.

Thirdly, the final clause in a complementizer cleft is always headed by *dat* ‘that’, but this is not the case for extraposed clausal subjects. For some of the adjectives that allow for *that*-clause extraposition we also find examples with *of*-clause extraposition (42) or VP extraposition (43). We conclude that cleft sentences with extraposed complementizer clauses are different from adjectives with extraposed clausal subjects.

- (42) Het is onduidelijk of de stijging iets te maken heeft
 it is unclear whether de increase something to make has
 met het nieuwe honderdje.
 with the new hundred-diminutive

It is unclear whether the increase has anything to do with the new one hundred guilder bill.

- (43) Het is goed om kritiek te krijgen van De Nederlandsche Bank
 it is good to critique to receive from The Dutch Bank
It is good to receive comments from The Dutch Bank

2.3.3 The intransitive analysis

Complementizer clefts have to be distinguished from relativizer clefts and from other constructions with extraposed complementizer clauses. In this section we analyze them as a separate construction, based on the intransitive (existential) copula.

The complementizer cleft consists of the copula and three components: the pronoun *het*, the c-focus and the final clause. We argued in section 2.3.1 that *het* is the expletive subject in this construction, among other things because it cannot be replaced by any other nominal.

The second constituent is the c-focus. It is unlikely to be the object of the copula, because it cannot be an NP. But if it is not OBJ, then what is it? Transformationalists analyzed it as a phrase that is moved out of the final clause (Pinkham and Hankamer, 1975; Emonds, 1976). They thus derived complementizer clefts (35-a) from canonical sentences (35-d). Pollard and Sag (1994) implemented the same idea in a non-transformational way. They assume a special lexical entry for *be* for clefts. This lexical entry has three elements on the subcategorization list: *het*, an XP and a complementizer clause with that XP on slash.

If the c-focus originates in the final clause, we expect to find “traces” of it in the clause. If the extracted material is an adjunct, it is impossible to tell whether something is missing in the clause. But if it is an argument, then we should find an embedded verb which lacks one of its argument. We do indeed find examples like (44), where the c-focus realizes an argument function of the embedded verb.

- (44) a. Het is van de onwetendheid van de mensen dat ik het moet
 it is of the ignorance of the people that I it must
 hebben.
 have
It is on the ignorance of the people that I depend.
 b. Het is in deze jeugdlektuur dat het duivelse en helse
 it is in this child-literature that the devilish and hellish

zich voor het laatst manifesteren.
 self for the last manifest
*It is in this children's literature that devilish and hellish things
 last manifest themselves.*

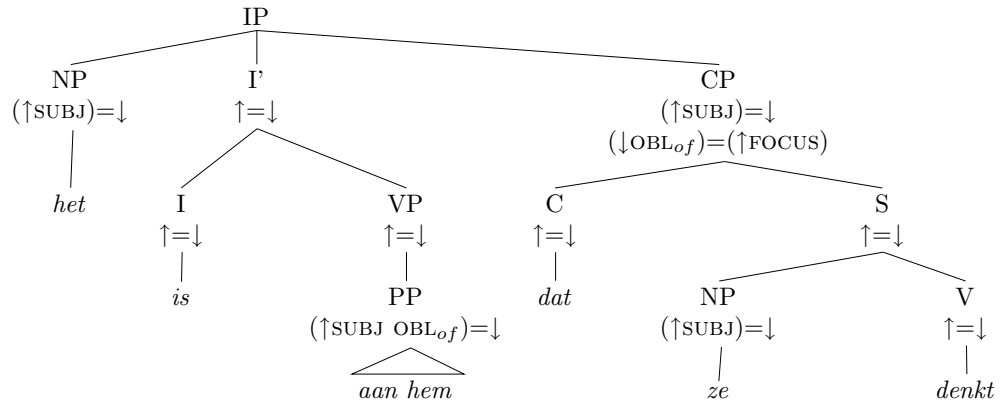
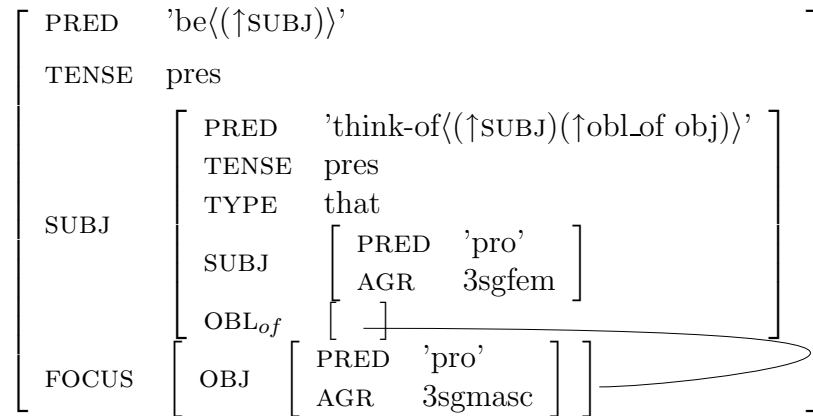
How does this clause fit in the argument structure? The complementizer clause is not in a canonical argument position, but in a sentence final position for extraposed constituents. It is not the object of the transitive copula, because the transitive copula needs two referential, non-expletive arguments (and we already showed that the c-focus is an extracted part of the clause and the subject is expletive). This is illustrated in (45). Here we repeated (35-a), but we undid the extraction of the c-focus out of the clause, so that it is a regular *that*-clause object. The sentence is syntactically marked, but even if one judges the sentence well formed (e.g. as an answer to the question what the main point of his lecture was), its meaning is different from the meaning of the cleft sentence, because the pronoun is interpreted referentially instead of as an expletive.

- (45) ?Het was dat het eerste vrije vakverbond onder het communisme
 it was that the first free union under the communism
 in Polen werd opgericht.
 in Poland was founded

Alternatively, we analyze the copula in the complementizer clause cleft as the intransitive copula. *Het* is in subject position and maps onto the SUBJ f-structure as dictated by the word order rules for Dutch. It does not contribute anything to the f-structure besides third person singular agreement values, because it is an expletive subject. The complementizer clause (with extraposed c-focus) is then mapped to the same SUBJ slot. This does not lead to a clash with *het*, because it unifies with the only features of the pronoun, the AGR features. Like the expletive pronoun, the complementizer clause is always third person singular, as illustrated in sentences with CPs in canonical subject position (46).

- (46) Dat we gewonnen hebben is nog niet zeker
 that we won have is still not certain
 It is not certain yet that we have won

This gives us a total of three c-structure nodes associated with the SUBJ f-structure slot: *het*, the complementizer clause and the clefted element, which is extracted from the CP. An example c-structure is shown in fig. 2.4. The corresponding f-structure is in fig. 2.5.

Figure 2.4: c-structure for *Het is aan hem dat ze denkt*Figure 2.5: f-structure for *Het is aan hem dat ze denkt*

2.3.4 Formalization

The c-structure rules for intransitive clefts are given in figure 2.6.¹⁸ The rules also account for sentences with embedded “gaps” in the complementizer phrase, as in the constructed example (47). The clefted element is situated in the canonical object position inside the VP. This is in contrast with analyses that assume the clefted element and the final clause to be one constituent Merchant (1998); Rizzi (1997). However, this assumption does not hold for Dutch clefts, since the verb cluster obligatorily follows the clefted element and thus separates the two phrases, as in the constructed example below (48). Furthermore, the NP plus *that*-clause cannot be topicalized, as one would expect if they formed a constituent.

- (47) Het was in Polen dat ze dacht dat het eerste vrije
 it was in Poland that she thought that the first free
 vakverbond onder het communisme was opgericht.
 union under the communism was founded
*It was in Poland that she thought the first free union under commun-
 ism was founded*
- (48) Het moet in Polen geweest zijn dat het eerste vrije vakverbond
 it must in Poland been be that the first free union
 onder het communisme was opgericht.
 under the communism was founded
*It must have been in Poland that the first free union under commun-
 ism was founded*

The rules do not specify the NP in the canonical subject position. This is not necessary, because the expletive *het* is the only NP that would not lead to a clash in that position: every other NP has a PRED, which cannot possibly unify with the PRED of the complementizer clause because of functional uniqueness. We did not specify the argument function of the c-focus either, which means that the complementizer clause has to be instantiated to determine the syntactic function of the c-focus (or the coherence principle is violated).¹⁹

¹⁸The constraint $(\downarrow\text{CPATH})=(\uparrow\text{FOCUS})$ expands to $(\downarrow\text{XCOMP}*\text{OBL}_\theta)=(\uparrow\text{FOCUS}) \vee (\downarrow\text{XCOMP}*\text{OBL}_\theta)\ni(\uparrow\text{FOCUS})$

¹⁹The definition of CleftP should be expanded to allow for the restricted set of nominals that appears in intransitive clefts. As we have not identified the constraints on the occurrence of nominals in this construction yet, and generally allowing NPs in this construction would lead to massive overgeneration, we decided to leave out NPs for now. Additionally, for speakers who judge examples with an adjectival c-focus (32-b) grammatical, CleftP and CPath should be expanded accordingly.

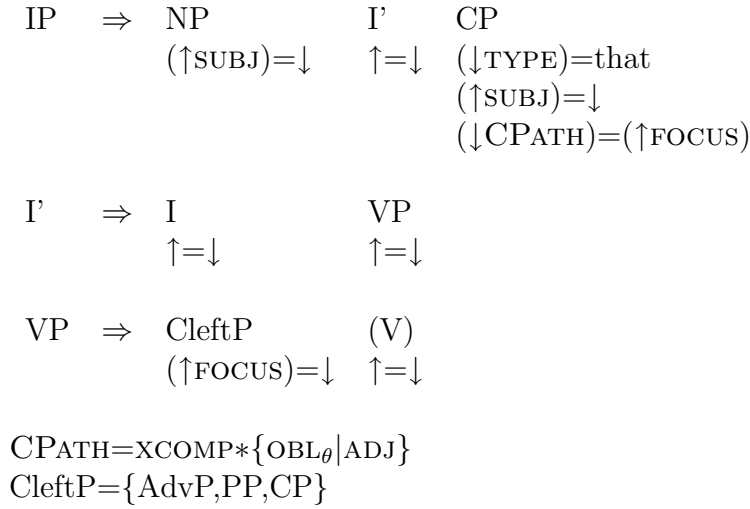


Figure 2.6: C-structure rules for non-nominal clefts

In section 2.3.1 we stated that the closest paraphrase of the intransitive cleft (35-a) is the simplex sentence (35-d). This is in line with the analysis presented in this section, which specifies the meaning of the cleft sentence to be the existential assertion of the simplex sentence (with focus on one particular constituent). Furthermore, the observation that transitive, but not intransitive clefts exist with other copulas than *zijn* follows automatically from our analysis: while the transitive copula *zijn* forms a natural class with other transitive copulas such as *lijken* and *blijken*, the latter do not have the intransitive, existential use that *zijn* has.

Note that the intransitive analysis would be inappropriate for the *it*-clefts with a relative clause, which we analyzed as transitive copular sentences. The intransitive analysis depends on the subject being expletive and we showed that this is not the case in relative clause clefts. Furthermore, the relative clause cannot independently function as an argument; it always needs an antecedent (unless it is a free relative). The two distinct analyses are furthermore motivated by the different semantics, informally illustrated by the different paraphrases. The two have in common that given information is extraposed to focus new information.

2.4 Conclusion

We have accounted for the syntactic differences between various realizations of the cleft construction by splitting up the data into two types of *it*-clefts: those with a final relative clause, and those with a final complementizer clause. We argued that the latter is a distinct construction, different both from the NP cleft and other constructions with superficially similar structures.

We analyzed the *it*-cleft with a nominal c-focus and a relative clause as an instance of the transitive copula and those with a complementizer clause as an instance of the intransitive verb *zijn* ‘to be’. We argued for a lexical entry for *het* ‘it’, *dat* ‘that’ and *dit* ‘this’ with optional agreement features. These account for the apparent lack of agreement in copular sentences with two NP arguments. Since the transitive cleft is an instance of such a sentence, we also accounted for the subject verb agreement pattern in this type of cleft. In addition, we accounted for the agreement on the embedded verb and agreement between the relative pronoun and the antecedent by analyzing it as a modifier of the TOPIC. This also explains the often observed givenness constraint in clefts: all the information in the clause has to be given or background information.

In the second type of cleft, the intransitive cleft, all phrases are associated with the subject function of the verb *zijn*: the c-focus is analyzed as an extracted constituent of the complementizer clause and both the CP and the expletive pronoun in subject position are unified with the subject function.

We did not have to stipulate construction-specific lexical entries for the copula, as clefts were analyzed as instances of the regular transitive and intransitive uses of *zijn* ‘to be’. All cleft-specific information was specified on the added, cleft-specific components in the c-structure rules. Both types of cleft involve extraposition of given information in order to focus new information.

We did not discuss the properties of the NP complement of the copula in the transitive cleft. It is clearly different from regular objects, for example in that it cannot passivize. But is different from predicative complements too, first of all in that it doesn’t have to be predicative: proper names are allowed too. In fact, we saw that purely predicative (bare) nominals are even excluded. A more precise definition of the constraints on the nominal complement and their characteristics was left for future research.

The fact that only pronominal c-foci appear in subject position was attributed to a more general phenomenon that pronouns have a strong preference for the subject function, much stronger than other syntactic categories. We expect that with further investigation along the lines of the research described

in the next chapter, a more detailed account for this particular distribution could be given.

Chapter 3

Dative Alternations

In this chapter we investigate the different realizations of the Dutch dative alternation and the factors that influence this construction. For English, general linearization constraints have been claimed to influence the realization of ditransitive verbs. But in contrast to English, the syntactic category of the recipient argument and the order of the arguments may vary independently in Dutch. We thus expect that those linearization constraints influence only the order of the arguments, not the choice for a dative NP or a dative PP. This hypothesis is tested on the basis of quantitative data from automatically annotated corpora and focusing on pronominality and weight constraints. On the other hand we expect lexical preferences to target only the syntactic category of the recipient argument, not the relative order of the arguments. This is tested using the same techniques. The factors shown to influence the dative alternation are formalized in the framework of Optimality Theory. Finally, we provide additional evidence for our analysis of argument order variation based on a second construction, the *Accusativus cum Infinitivo*.

3.1 Introduction

The dative alternation is a very well-studied linguistic phenomenon and it is by no means specific to Dutch: much work has been done on the dative alternation in English, illustrated in (1) (Givón, 1984; Pinker, 1989; Levin, 1993; Goldberg, 1995; Wasow, 2002; Krifka, 2001; Levin and Rappaport Hovav, 2002; Bresnan and Nikitina, 2003; Bresnan et al., 2005, for example). Many different factors have been claimed to influence the dative alternation in English, ranging from general linearization principles, to semantics, lexical preferences and person effects.

- (1) a. Jo gave the student the book

- b. Jo gave the book to the student

In Dutch, the alternation is a bit more complex. In many aspects, the syntactic construction resembles the English situation: although we adopt the by now established term ‘dative alternation’ to refer to the different realizations of verbs with a recipient and a theme argument, no actual dative casemarking is involved, just as in English. Both NP and PP recipients take the non-subject form, which can only be distinguished from the subject form if the recipient is pronominal. However, in addition to the canonical double object construction and the dative PP construction in (2-a) and (2-b), Dutch has two more variants: both the double object construction and the PP construction occur with non-canonical word orders. In (2-c) we find the direct object shifted in fronted of the indirect object. In (2-d) we see that the ‘dative’ PP is shifted and precedes the direct object. Both variations violate the rule for canonical argument order $\text{SUBJ} < \text{OBJ2} < \text{OBJ1} < \text{OBL}, \text{XCOMP}$ for Dutch.¹

- (2) a. Vervolgens gaf hij mij geel.
 afterwards gave he me yellow
And then he gave me a yellow card.
- b. Vervolgens gaf hij geel aan de speler.
 afterwards gave he yellow to the player
And then he gave a yellow card to the player.
- c. Vervolgens gaf hij het mij.
 afterwards gave he it me
And then he gave it to me.
- d. Vervolgens gaf hij aan die speler een eerste officiële
 afterwards gave he to that player a first official
 waarschuwing.
 warning
And then he gave a first official warning to that player.

The existence of the non-canonical variants in Dutch teases apart two distinctions that are merged together in the English situation: where English has two variants which differ with respect to both the syntactic category and the

¹Throughout this chapter, we will use both ‘indirect object’ and OBJ2 to refer to the grammatical role to which the recipient argument is mapped. This is contrary to much work in Lexical Mapping Theory on English, where the recipient is assumed to map to OBJ1 in the double object construction. There is reason to assume that English and Dutch differ in this respect, e.g. Dutch does not generally allow the recipient of a ditransitive verb to be mapped onto the subject function in passive sentences. It does allow passive sentences with theme subjects and recipient objects.

order of the arguments,² Dutch has two separate alternations: one between NP and PP recipients and one between the canonical and non-canonical word order. This results in a total of four different realizations for ditransitive verbs, which give us the opportunity to study the two alternations separately. One could expect, then, that the factors that have been claimed to influence the dative alternation in English, now fall in two categories. A first category which influences the syntactic category of the recipient argument, for example a lexical preference of the verb for an NP or a PP recipient, and a second category which contains general constraints on argument ordering. In this chapter we investigate whether this expectation is borne out.

This chapter is structured as follows. We start with a brief discussion of the literature on the dative alternation in English, focusing on the different factors that have been claimed to influence the alternations (section 3.2). We then investigate to what extent these factors influence the Dutch dative alternation, and more specifically whether they influence the argument order alternations (sections 3.4 and 3.5) or the NP/PP alternation (section 3.6). We incorporate our findings in the framework of Optimality Theory (OT). The advantage of modeling the influencing factors in OT is that it allows for violable constraints and constraint interaction. As we will see, there are many constraints that are important for the dative construction, but most of them can be overridden. The constraint on canonical word order, for example, can be overridden in order to avoid a right aligned pronoun. This is difficult to model in other grammatical frameworks. We will see, however, that the dative alternation poses challenges for OT as well. Section 3.8 shows how the argument ordering constraints that we identified for the dative alternation can be applied to account for another case of argument reordering in Dutch, the AcI construction. Finally, in 3.9 we summarize our findings and conclude.

3.2 Previous Work

The literature on the dative alternation can be divided in two categories: analyses that focus on the order of the two complements, and analyses that focus on the category of the recipient argument. We start with a brief overview of the first category and continue in section 3.2.2 with analyses of the second category.

²Even for English, the inverted word order is occasionally found, mostly in biblical texts (i). We will not go into these exceptional examples here.

(i) ...and I will give to him a white stone, and on the stone a new name written, which no one knows but he that receives it. (Revelations 2:17)

3.2.1 Linearization Constraints

The issue of constituent ordering has been discussed extensively in the linguistic literature. This section does not aim to be an exhaustive discussion of the complete literature on this topic, but merely an overview of various approaches with some pointers to work within that approach.

One approach to linearization can build on a particularly long tradition: Behaghel already observed in 1909 that long and complex phrases tend to follow lighter material (Behaghel, 1909/10). Since then, this observation has been applied to various ordering phenomena, including the dative alternation Arnold et al. (2000); Wasow (2002); Hawkins (1994); Erteschik-Shir (1979). The general idea is that although the double NP construction is generally favored, heavy recipients may be realized as (right-aligned) PPs in order to avoid a violation of the general principle on word order that says that heavy constituents align right. The influence of weight on word order is widely accepted and rephrased for other languages, including German (Uszkoreit, 1987) and Dutch. The Dutch reference grammar (Haeseryn et al., 1997) formulates this influence in the Complexity Principle, which states that light constituents precede heavy constituents.

Almost as widespread is the idea that information structure influences word order. Gundel (1988) and Prince (1992) showed that the contrast between old and new information is in this respect similar to the contrast between light and heavy material: new information tends to follow old, topic information. In addition, Arnold et al. (2000) showed that although weight and givenness are not independent of each other, they do both have a distinct effect on word order. Haeseryn et al. (1997) rephrase this constraint as the Left-Right Principle, which is claimed to be the main principle responsible for ordering in the Dutch midfield. Uszkoreit (1987) did not formulate a similar principle for German, but instead broke up the influence of information structure in two separate constraints: pronouns precede full NPs and definites precede indefinites. As pronouns are by definition old information but NPs not, and as definites are often old, but indefinites usually new, it is not hard to see that the Left-Right Principle and Uszkoreit's constraints lead to similar results.

German and Dutch, though allowing for word order variation, do have a clear canonical word order that is generally preferred. This may be stated as a simple constraint, e.g. “subject precedes indirect object which precedes direct object” (Uszkoreit, 1987). The Dutch reference grammar has formulated the Inherence Principle. This principle says that whatever is more closely connected to the main verb, should be realized closer to the second pole (i.e. the right edge of the midfield, which is defined to be the position of the verb

cluster). As the theme is generally considered more closely connected to the verb than the recipient, this principle favors the direct object being realized closest to the second pole. This is in line with the canonical word order in double object constructions, but clashes with canonical word order in dative PP constructions. The Inherence Principle is related to the “natural constituent structure” (Vennemann, 1973). This principle, attributed to Renate Bartsch, states among other things that the closeness of constituents in the surface string reflects the hierarchical dependencies between them. More recently, Wasow (2002) formulated the concept of semantic connectedness. Wasow states that constituents that are closely connected semantically are very likely to appear in adjacent positions. He furthermore argues that this constraint may override the other constraints (weight, information structure) on word order.

Furthermore, ambiguity avoidance may influence linearization. Wasow (2002) tested its influence and concluded that the aim to avoid ambiguity has at most a minor effect on the ordering of constituents. In this paper we will see some data which may point to such an influence, but conclusive evidence is yet to be found.

In addition to these general linearization constraints, it is well known (e.g. Lenerz (1992)) that pronouns behave differently from lexical NPs with respect to word order. We already saw that pronouns generally precede full NPs (Uszkoreit, 1987). Two constructions in which this principle plays a crucial role are Wackernagel Movement (WM) in German and Object Shift (OS) in Scandinavian languages. WM involves the obligatory shift of certain types of pronouns in German to a left-peripheral position, preceding all other nominal arguments (except subjects, which may precede the pronoun). Müller (2001) accounts for WM making use of a personal pronoun scale. As we will see, a similar scale exists in Dutch. Several Scandinavian languages show a similar flexibility in the midfield, allowing reduced pronouns to ‘shift’ leftwards, as long as they do not precede the verb they are an argument of (Diderichsen, 1946; Börjars et al., 2003). For a unified account of WM and OS, see Thráinsson (2001).

A radically different approach on word order is taken by Reinhart (1996), who formulated a focus-driven approach to word order. She argues that the sentence focus depends on the position of the main stress: the focus of IP is a(ny) constituent containing the main stress of IP. Usually, main stress falls on the right edge of the middle field in Dutch. If the focus of IP is a constituent that does *not* contain the rightmost phrase in the middle field, focus and stress do not coincide naturally. There are two ways to fix this: either the stress shifts to another position or the order of the constituents is changed through scrambling. Reinhart claims that scrambling is more

economical than stress shift and therefore the preferred strategy for stress (and thus focus) assignment.

The non-canonical versions of the double object construction and the dative PP-construction could be regarded as a form of scrambling and one could argue that even the NP/PP alternation could be explained in the same manner, as the canonical argument order for the dative PP and the double object construction do differ. Such a focus based approach would nicely explain why the phonologically weak pronoun *het* almost always shifts. After all, this pronoun is never stressed. It would also account for the fact that we find very few emphasized forms of the pronouns (e.g. *hijzelf*, ‘he himself’) in shifted position: the emphasized forms are focused and thus should stay at the right edge of the middle field in the canonical stress position. However, the data in (3) (grammaticality judgments hers), on which Reinhart (1996) bases her theory, are controversial, to say the least.

- (3) a. *Ik heb de krant nog niet gelezen, maar ik heb het boek
 I have the newspaper yet not read but I have the book
 al wel gelezen.
 already indeed read
- b. Ik heb nog niet de krant gelezen, maar ik heb al
 I have yet not the newspaper read but I have already
 wel het boek gelezen.
 indeed the book read
I haven’t read the newspaper yet, but I did read the book already.

In any case, focus cannot be the full explanation for the non-canonical argument order: the alternation between canonical (4-a) and non-canonical (4-b) orderings persists even if both arguments are phonologically weak pronouns and therefore necessarily unstressed (4). In what follows, we do not discuss the role of stress and focus in the dative alternation, but instead focus on lexical and syntactic features.

- (4) a. “Blijf liggen”, zei hij, “ik geef je het wel.”
 stay lying said he I give you_{weak} it_{weak} so
“Stay down”, he said, “I’ll hand it to you”.
- b. Ja ’k zal ’t ’m zeggen.
 yes I will it_{weak} him_{weak} say
Yes, I’ll tell him that.

We have discussed four factors which may influence the order of the arguments in the dative alternation: canonical word order, weight, information

structure/focus, pronominality and ambiguity avoidance. In the research reported on below, information structure and ambiguity avoidance are briefly mentioned when applicable, but the main focus is on canonical word order, weight and pronominality.

3.2.2 The NP/PP alternation in English

As argument order and recipient realization are inseparable in the dative alternation in English, analyses of the construction have employed both the difference in ordering and the difference in syntactic category. In the previous subsection, we highlighted some approaches that focus on general ordering principles. We now turn to some analyses of the dative alternation that focus on the NP/PP alternation specifically.

Levin (1993) argues that verbs can be classified according to their meaning. Within these classes, all verbs are assumed to show similar syntactic behavior, but across classes, their behavior may vary. For instance, *send verbs* (e.g. ‘hand’, ‘mail’) and *transfer of message verbs* (e.g. ‘ask’, ‘read’) are assumed to alternate, while *manner of speaking verbs* (e.g. ‘whisper’, ‘shout’) and *bill verbs* (e.g. ‘bill’, ‘fine’) are not: these verbs supposedly only occur with the PP and NP alternant respectively. Lapata (1999) tested the empirical value of the semantic verb classes described by Levin (1993) with corpus based methods and concluded that there are statistically significant differences in the frequencies with which certain verbs occur in NP and PP ditransitive constructions. On the other hand, Bresnan and Nikitina (2003) convincingly showed that these are mere tendencies, rather than categorical differences.

Krifka (2001) and Pinker (1989), among others, take this one step further. Building on the work by Green (1974), they adopt a classification of verbs in verb classes, and argue that the reason that certain classes of verbs are incompatible with one of the two realizations is that the variants have different meanings and the verb meaning may be incompatible with one of them. The assumed meaning for the PP construction can be paraphrased as ‘x causes y to go to z’ and the meaning of the double NP construction as ‘x cause z to have y’. According to this line of explanation, there is no dative alternation proper: the double object construction and the PP construction are not alternative ways of expressing the same meaning, but they are realizations of different meanings. Bresnan and Nikitina (2003) provide examples of alternating dative syntax in contexts of repetition, which form a challenge for this approach. The view that the dative alternation reflects a difference in meaning contrasts with the widespread view that the verb means the same in both constructions, whether the two are related via syntactic transform-

ations (Larson, 1988; Aoun and Li, 1989) or different argument expressions of the same verb (Butt et al., 1997). Monosemic analyses of the dative alternation only explain the existence of the two structures, not the choice for one of them in a specific situation. They either predict free variation or require additional mechanisms such as general ordering or lexical constraints to determine the particular realization of a ditransitive verb.

Besides lexical preferences and semantic differences, there are also accounts of the NP/PP alternation (partly) based on structural constraints on the realization of the dative argument. Bresnan and Nikitina (2003) argued that local (1st or 2nd person) recipient arguments prefer to be realized as objects, not obliques, while nouns prefer the dative PP structure. This is formalized in the violable constraint conjunction $\text{HARMONY}(1,2): *NP_{\text{Noun}} \& *PP_{1,2\text{Person}}$. As a result of $\text{HARMONY}(1,2)$, local recipients will generally lead to double object constructions instead of dative PP constructions. Bresnan et al. (2005) furthermore report that the animacy of the recipient heavily influences its syntactic category (with inanimate recipients strongly preferring the dative PP structure).

This research will touch on the influence of person and animacy, but it will focus on verbal preferences for a PP or a dative NP structure. In English, these preferences will influence both the syntactic category of the recipient argument and the relative order of the theme and the recipient, because these two factors are inseparable. The hypothesis is that in Dutch, where syntactic category and order may vary fairly independently of each other, these lexical preferences influence only the choice for an NP or PP, not the argument ordering. This hypothesis is tested on the basis of frequency information from corpus data. The influence of these verbal preferences is then incorporated in our OT model of the dative alternation.

3.3 Preliminaries

In this section we describe the lexical resources that we used and the methodology that we applied for the corpus-based research described in the following sections of this chapter.

3.3.1 Resources and methodology

Corpora contain valuable information about the distribution of different realizations of the dative construction. A potential problem is that the structures we are interested in (the four alternants of the Dutch dative alternation) are

specific and complex syntactic structures, which cannot be retrieved from corpora on the basis of simple pattern recognition. Therefore, we used syntactically annotated and automatically parsed data in our corpus study. Two annotated corpora were used, the annotated part of the Corpus of Spoken Dutch (CGN, about 1M words (Levelt, 1998)) and the Alpino Treebank (the annotated CDB newspaper part of the Eindhoven Corpus, about 150K words (van der Beek et al., 2002a)), which are both annotated with dependency structures (Moortgat et al., 2001).

When the annotated corpora proved too small for statistically relevant results, we used a corpus of over 75M words of newspaper text (Twente Nieuwscorpus) that was automatically parsed by the Alpino parser (Bouma et al., 2001; van der Beek et al., 2002b). The parser outputs the same dependency structures as those used in the annotated corpora. With a 85.5% parsing accuracy (measured over the dependency relations), the quality of the annotation in the automatically parsed corpus is lower than the manually annotated corpora—although 100% accuracy is never reached, not even in hand-annotated corpora. The biggest problem with respect to the performance of the grammar is to make sure that the data is found, given that it is in the corpus. How do we know that the grammar will not systematically misparse certain double object constructions? The chances that the grammar does not recognize a dative construction are small. The hand-written Alpino grammar overgenerates, putting very few constraints on the argument order in double object constructions and always building a dative parse if possible. However, from all parses that are generated, only one is included in the corpus: the most probable parse is selected by a statistical Maximum Entropy model. We cannot look into this model to see what preferences it has deduced from manually annotated data. Here, we must rely on manual inspection of the data, to see if 1) the extracted candidates are true datives and 2) to see if any obvious constructions are missing. We did not find any signs of systematic errors in the data. As an extra precaution, we used queries that abstracted away from the argument order when extracting dative constructions from automatically parsed data (including both orders and sorting them manually), so that ordering errors by the parser did not influence the data collection.

The syntactically annotated corpora were queried using `dt_search` (Bouma and Kloosterman, 2002), a tool which allows us to query the treebank on dependency relations, syntactic category or part-of-speech and linear order. See chapter 1 for more general information on Alpino or `dt_search`.

We excluded from our search all instances of (in)direct object topicalization, all (wh)relativizer direct and indirect objects and all clausal objects such as that-clauses because in these sentences, the order of the arguments is

	NP NP _{unshift}	NP NP _{shift}	NP PP	PP NP	TOTAL
CGN	226	33	63	8	334
Alpino	122	7	43	10	182

Table 3.1: Distribution of the three alternants of the dative alternation Dutch manually annotated corpora.

determined by other factors. Passive sentences and instances of the *krijgen*-passive (the “*get*-passive”, a dative passive construction) were also excluded. The motivation for this is that the direct object (in the regular passive) or the indirect object (in the *krijgen*-passive) surfaces as the subject of the matrix clause, therefore the word order rules for subjects applies here. Finally, we excluded all instances of *er*-recipients. In these sentences, illustrated in examples (5), the recipient argument is third person, inanimate and singular and realized as a pronoun inside a PP. In these cases, so-called R-pronouns (*er/daar* ‘there’, *hier* ‘here’) are used instead of the regular third person neuter singular *het* ‘it’ and this pronoun is often fronted. The preposition stays in position, resulting in a split PP. The alignment of *er* is a characteristic of R-pronouns, not a characteristic of the dative construction.

- (5) Ik geef daar geen les aan.
 I give there no class to
I won't teach those.

3.4 Linearization: the Double Object Construction

In both the double object construction and the dative PP construction, the canonical word orders (NP NP_{unshift} and NP PP) are much more frequent than the shifted alternants (see fig. 3.1). Furthermore, the double object construction is much more frequent than the dative PP, and the canonical word orders (NP NP_{unshift} and NP PP) are much more frequent than the shifted alternants.

Although the general distribution is highly skewed, the chances of finding one of the less frequent realizations increases considerably under certain conditions. The following sections investigate what those conditions are. We first look at general linearization factors that influence the order of the objects in the double object construction.

	NP	NP _{unshift}	NP	NP _{shift}	NP PP	PP NP	TOTAL
CGN		143		33	57	3	247
Alpino		45		6	21	3	83

Table 3.2: Distribution of dative alternation realizations with one word themes in manually annotated corpora.

3.4.1 Pronominality

The general distribution of the dative alternation changes drastically if we restrict the object to one lexical item only (we do allow additional function words such as determiners). While the numbers for the shifted double object construction hardly changed in table 3.2, the numbers for the unshifted and PP variants dropped by 10-70%. This may look like a strong influence of weight on constituent ordering, but it is actually caused by the fact that a shifted direct object almost exclusively occurs with pronominal direct objects.

Of all shifted direct objects in our manually annotated data, only one example contained a shifted full NP (6). In this example, we find the archaic dative marking on the indirect object. We assume that it is this overt dative marking that makes available the freer word-order and that Direct Object Shift (DOS) is generally restricted to pronouns.³

- (6) [daar] heeft Paul Badura-Skoda het nieuwe pianoconcert van Frank
 there has Paul Badura-Skoda the new piano_concert of Frank
 Martin den muzikale volke voorgesteld.
 Martin the_{dat} musical_{dat} people_{dat} presented
there, Paul Badura-Skoda presented the Frank Martin's new piano concert to the musical people.

With its restriction to pronominal objects, the Dutch object shift resembles the Scandinavian Object Shift (OS) and Wackernagel Movement (WM) in

³However, Zwart (1997) presents examples that show that NP-DOS with definite NPs is not impossible:

- (i) dat Jan het boek Marie terug gegeven heeft.
 that Jan the book Marie back given has
that Jan gave the book back to Marie.

No examples of this kind were found in our corpora. We suspect the exceptional definite NP shift to be a focus effect and leave this and other effects of focus on word order for future research.

German. But there are several differences: Scandinavian OS and WM apply to both direct and indirect objects, and may involve shifting beyond the subject. As a result, we find (in)direct objects preceding the subject in those languages. In Dutch, indirect objects may shift in front of the subject in some cases, although this is rare in ditransitive sentences. Direct objects on the other hand never precede subjects unless via topicalization. Furthermore, the Scandinavian shift depends on the position on the verb. This is not the case in Dutch.

English also has a constraint on the distribution on pronouns in the dative alternation: *NP PRO (Bresnan and Nikitina, 2003; Erteschik-Shir, 1979; Collins, 1995). This constraint states that personal pronouns, but not demonstratives or indefinite pronouns, are avoided when following full NPs if both are objects. The Dutch situation differs from English in that shift also occurs with demonstrative pronouns as well as personal pronouns, and that *het* ‘it’ not only precedes NPs but also other personal pronouns.

Not all direct object pronouns shift always. While the pronoun *het* ‘it’ usually shifts irrespective of the category of the indirect object, most other personal pronouns and the demonstratives shift if the indirect object is a full NP (7-a)⁴, but stay in their canonical position if the indirect object is a personal pronoun (7-b). First and second person pronouns do not shift. Not only are there no shifted local pronouns in the corpus, but made up examples of local pronoun DOS also lead to ungrammaticality under the intended reading (indicated with a %; example (8-b) is grammatical under the reading without DOS, i.e. the reading with a recipient *jou* ‘you’).

- (7) a. Ja, vertel dat de buurvrouw maar.
 yes tell that the neighbour DISC-PART
 Yes, go ahead, tell it to the neighbour.
 b. Heeft hij je dat niet verteld?
 Has he you that not told
 Didn't he tell you that?
 c. 'K zal 't hem zeggen.
 I will it him tell
 I'll tell it to him
- (8) a. De student wijst 'm de docent aan.
 the student points him the teacher at
 The student points him out to the teacher.
 b. %De student wijst jou de docent aan.
 the student points jou the teacher at

⁴DISC-PART indicates a discourse particle.

Shifted		Canonical	
542	het (it)	372	dat (that)
45	dat (that)	83	dit (this)
21	't (it _{reduced})	51	het (it)
19	ze (them)	28	die (that)
7	dit (this)	24	hem (him/it)
4	u ⁵ (you _{honorific})	14	zich (himself/herself)
4	hem (him/it)	8	hetzelfde (it _{same})
4	die (that)	4	me (me)

Table 3.3: Direct object pronouns in constructions with two pronominal objects

The student points you out to the teacher.

Table 3.3 shows the most frequent direct object pronouns in double object constructions where both arguments are pronominal. The data are based on the automatically parsed TwNC corpus. The frequency lists confirm the intuition that *het* shifts while demonstratives usually do not shift in front of another pronoun. Importantly, the table shows that the distinctions are not categorical: we do find *het* (it) in the canonical object position, although ten times less frequently than in the shifted position. Manual inspection showed that these unshifted *het* objects are not (all) the result of parse errors. The one place where we would not expect any variation is with the local pronouns, as even made up examples were ungrammatical. It is therefore surprising that we do find four occurrences of *u* (you_{honorific}). Further inspection showed that these are the result of parse errors, however.

The pattern of pronouns preceding full NPs, and *het* ‘it’ preceding both pronouns (demonstratives, animate pronouns) and full NPs resembles the differentiation of pronouns proposed for German by Müller (2001) to account for Wackernagel movement (9). He argues that if a certain pronoun ‘moves’ under certain conditions, then all weaker pronouns do, too. In the Dutch dative alternation, there are two different conditions: either the indirect object is a full NP or it is a pronoun. In the first case, anything weaker than a local pronoun shifts. In the second case, usually only *het* ‘it’ shifts.

- (9) Pron_{strong} > Pron_{unstressed} > Pron_{weak} > Pron_{reduced} > (Pron_{clitic})
 +stress +anim -anim es ‘it’ ’s

⁵All occurrences of the local pronoun *u* result from parse errors.

We describe this pattern with a set of linearization constraints. All constraints take the form $\text{PRO}_x < y$, where x indicates a type of pronoun and y indicates some other type of argument nominal.⁶ The pronoun scale differentiates between animate and inanimate pronouns. Although the dative alternation does not provide evidence for this differentiation, we formulated separate constraints. We thus predict to find differences in the distributions of the two pronouns. In section 3.8 where we treat the Dutch AcI-construction, we will see that this expectation is borne out.

- (10) $\text{PRO}_{it} < \text{NP}/\text{PRO}$: the pronoun *het* precedes NPs and pronouns in the midfield.

$\text{PRO}_{3rd/inanim} < \text{NP}$: inanimate personal and demonstrative pronouns precede full NP arguments in the midfield.

$\text{PRO}_{3rd/anim} < \text{NP}$: animate personal and demonstrative pronouns precede full NP arguments in the midfield.

$\text{PRO}_{local} < \text{NP}$: local pronouns precede full NP arguments in the midfield.

The constraints on the linearization of pronouns are in competition with the constraint on canonical word order (11): a pronominal direct object violates $\text{PRO} < \text{NP}$ if it follows the NP indirect object, and it violates the canonical word order constraint if it precedes the indirect object. From the non-canonical example sentences (7) we can conclude that some of the constraints on pronoun linearization outrank the canonical word order constraint, but the ungrammaticality of (8) shows that not all pronoun constraints do. The interaction of the constraints is illustrated in tableau 3.4. We merged the linearization constraint on animate and inanimate personal/demonstrative pronouns for clarity, and included references to (made up) example sentences. It is important to note that in these examples, a star indicates a candidate that is *categorically ungrammatical*, while a question mark indicates a *significantly less frequent* candidate.

- (11) $\text{O2} < \text{O1} < \text{OBL}$: OBJ2 precedes OBJ1 precedes OBL

⁶We merged the two constraints $\text{PRO}_{it} < \text{NP}$ and $\text{PRO}_{it} < \text{PRO}$ for clarity.







Input: <i>gives</i> (<SUBJ><OBJ1><OBJ2>)		Pro_{it} <NP/Pro	Pro_{3rd} <NP	$O2 < O1 < OBL$	Pro_{loc} <NP	
OBJ1='the book' OBJ2='de student'	 NP NP _{unshift} NP NP _{shift}			*!		(12-a) (12-b)
OBJ1='it' OBJ2='de student'	 NP NP _{unshift} NP NP _{shift}	*!		*		(12-c) (12-d)
OBJ1='it' OBJ2='him'	 NP NP _{unshift} NP NP _{shift}	*!		*		(12-e) (12-f)
OBJ1='that' OBJ2='the student'	 NP NP _{unshift} NP NP _{shift}		*!		*	(12-g) (12-h)
OBJ1='that' OBJ2='him'	 NP NP _{unshift} NP NP _{shift}			*!		(12-i) (12-j)
OBJ1='you' OBJ2='the student'	 NP NP _{unshift} NP NP _{shift}			*!	*	(12-k) (12-l)

Table 3.4: Shifted vs. canonical double object constructions

- (12)
- a. Bo geeft de student het boek.
Bo gives the student the book
 - b. *Bo geeft het boek de student.
Bo gives the book the student
 - c. ?Bo geeft de student het.
Bo gives the student it
 - d. Bo geeft het de student.
Bo gives it the student
 - e. ?Bo geeft hem het.
Bo gives him it
 - f. Bo geeft het hem.
Bo gives it him
 - g. ?Bo geeft de student dat.
Bo gives the student that
 - h. Bo geeft dat de student.
Bo gives that the student
 - i. Bo geeft hem dat.
Bo gives him that
 - j. ?Bo geeft dat hem.
Bo gives that him
 - k. Bo raadt de student_{io} jou_{do} aan.
Bo recommends the student_{io} you_{do} PART
 - l. *Bo raadt jou_{io} de student_{do} aan.
Bo recommends you_{io} the student_{do} PART

The tableau shows how the constraints interact to account for the most frequent patterns in various types of sentences. Pronominal direct objects will shift if the indirect object is a full NP, in order to avoid a violation of the constraint on the linearization of pronouns, which is higher ranked than $O2 < O1 < OBL$. If both objects are pronominal (but not *het* or local), $O2 < O1 < OBL$ is the highest ranked constraint that is violated by one argument order but not the other and thus determines the optimal candidate: OBJ2 precedes OBJ1. *Het*, on the other hand, will shift no matter what the category of the indirect object is, because a violation of $PRO_{it} < NP/PRO$ is worse than any other right aligned pronoun or a non-canonical word order. $PRO_{local} < NP$ is outranked by $O2 < O1 < OBL$, preventing local pronouns from shifting.

Our findings contradict the claim in Zwart (1996) that only what he calls ‘reduced’⁷ direct object pronouns can shift: the demonstratives were

⁷Zwart’s notion of ‘reduced’ does not correspond to weak pronouns in the personal

among the most frequently shifted pronouns and we also found non-reduced examples of third person pronouns. We do see a tendency, though, of the third person reduced pronouns *'m* (him, it) and *ze* (them) to group with *het* if the antecedent is inanimate. In this case, they tend to shift, even if the indirect object is a pronoun. We do not have enough data for a quantitative evaluation of this intuition, but integration of it in our model is straightforward if it proves correct: they would be grouped together with *het*. In any case, it does not seem very likely that *all* reduced pronouns shift: objects that take the form of reduced local pronouns (*me* ‘me’, *je* ‘you’) were never found preceding the indirect object.

3.4.2 Gradient patterns

The tableau shows the interaction for the most frequent pattern for each combination of direct object and indirect object, but we have seen that the preferences are not categorical. Classic OT cannot account for these less frequent patterns, as it assumes a strict ranking of constraints (13). Given this strict ranking, C_3 can never dominate C_1 or C_2 , and alternative outcomes are ruled out. Variation is only possible in classic OT if the constraints on which two candidates differ are equally strong (i.e. in a stratum, see also table 1.5 in section 1.3.2). In this case, free variation is predicted. However, we have seen that one pattern may be much more frequent than other patterns: we do not find the fifty-fifty distribution that we would expect if it were free variation.

- (13) Strict constraint ranking in classic OT
 $C_1 \gg C_2 \gg C_3$

We would need the stochastic OT implementation of Boersma and Hayes (2001) to account for the alternative patterns (see chapter 1 for a brief introduction in Stochastic OT). Boersma and Hayes assume that the constraints are ranked on a linear scale, with higher values corresponding to stronger constraints, and that they are evaluated stochastically. Whenever a candidate set is evaluated, the exact position of a constraint on the scale is determined. This exact numeric value depends on its ranking, but is perturbed by a random variable, which models the noise in the system. For two constraints $C_1 \gg C_2$ that are ranked closely together, it is possible that because of this noise, the actual selection point of C_2 is sometimes higher than for C_1 , leading to an alternative ranking and a different optimal output.

pronoun hierarchy (Müller, 2001) above. Zwart (1996) distinguishes between different forms of the pronoun, not between properties such as person or animacy.

Returning to argument order alternations, we find both categorical distinctions and gradient patterns in the DOS. As local pronouns never shift, the distance between the canonical word order constraint and the constraint on local pronouns must be large enough for the reversed order to be practically impossible. On the other hand, we do find infrequent occurrences of the canonical word order with a 3rd person direct object pronoun and an NP indirect object (14): on a total of 139 occurrences of this combination in the TwNC corpus, 19 were in the canonical word order.

- (14) Als je de patiënt dat kunt besparen, moet je dat doen.
 if you the patient that can save must you that do
If can can spare this to the patient, you should.

Thus, the distance between the canonical word order constraint and $\text{PRON}_{3rd} < \text{NP}$ must be smaller, allowing for a chance that canonical word order outranks linearization constraints in some evaluation.⁸

3.4.3 Weight

Under our present analysis, the DOS is solely driven by the syntactic category of the objects: NPs, personal or demonstrative pronouns or *het*. But pronominality is not independent of syntactic weight: pronouns are the lightest possible NPs. Thus, the pronominal DOS is in line with the Complexity Principle and Uszkoreit's weight principle. But we did not differentiate between heavy NP recipients and light NP recipients, although the weight principle would predict the former to allow DOS more easily than the latter. Table 3.5 lists the average weight of the direct and indirect object in all four variants of the dative alternation, as well as the OBJ1/OBJ2 weight ratios.⁹ The weight is simply expressed as the number of words (discarding the preposition in the PPs). More sophisticated definitions of weight could be applied, such as number of nodes in a syntactic tree. However, it has been shown that different formulations of syntactic weight all lead to similar results (Wasow, 2002; Szmrecsányi, 2004). We see that the average weight of the indirect object in shifted double NP constructions (1.09 and 1.71) is *lower* than in the canonical double object construction (1.40 and 2.43), con-

⁸We also found non-canonical examples of demonstrative pronouns with personal pronouns indirect objects: 18 of 506 occurrences. To account for this word order, we need to assume a constraint $\text{PRO}_{3rd} < \text{PRO}$, similar as for the pronoun *het* (see also footnote 6), which is usually outranked by the constraint on canonical word order. With STOT, there will be some (small) chance on this constraint outranking canonical word order.

⁹We controlled for extraposition by only including sentences in which recipient and them were followed by the verbal cluster which indicates the right edge of the middle field.

				OBJ1	OBJ2	OBJ1/OBJ2
CGN	NP	NP _{unshift}	(N=231)	3.75	1.40	2.68
Alpino	NP	NP _{unshift}	(N=123)	5.87	2.43	2.42
CGN	NP	NP _{shift}	(N=33)	1.03	1.09	0.94
Alpino	NP	NP _{shift}	(N=7)	1.71	1.71	1.00
CGN	NP	PP	(N=56)	2.02	1.93	1.05
Alpino	NP	PP	(N=17)	2.53	2.29	1.10
CGN	PP	NP	(N=7)	4.71	2.71	1.73
Alpino	PP	NP	(N=8)	3.50	3.25	1.08

Table 3.5: Average weight per grammatical role in number of words.

trary to what the Complexity Principle would predict. If we control for a pronominality effect and exclude pronominal recipients, the average recipient weight increases and the differences between the two alternants get smaller. But with averages of 2.35 (Alpino) and 1.68 (CGN) for the canonical double object construction and 3.18 (Alpino) and 1.93 (CGN) for the canonical dative PP construction, the PPs are still heavier than the NPs. We conclude that syntactic weight does not have the expected influence on the DOS. This is surprising, as weight is generally assumed to influence linearization via the principle ‘light precedes heavy’.

We can conclude that the argument order in the double object construction is influenced by one important constraint on word order in general, namely the principle that states that pronouns precede full NPs, which in turn is related to the principle that old information precedes new information. The other important linearization principle, light constituents precede heavy constituents, could not be shown to influence the ordering of the two objects in the expected way.

3.5 Linearization: Dative PP Shift

We now turn to the second ordering alternation in the dative construction in Dutch: the argument order alternation in the recipient PP construction. This dative PP is most often realized in its canonical position following the direct object (15-a), but can also be found preceding the direct object (15-b), where it violates the principle of canonical word order.

Dataset			OBJ1	OBJ2	OBJ1/OBJ2
CGN	NP PP	(N=63)	1.62	2.57	0.63
Alpino	NP PP	(N=43)	3.70	5.21	0.71
CGN	PP NP	(N=8)	5.63	1.63	3.45
Alpino	PP NP	(N=10)	3.00	3.30	1.45

Table 3.6: Average number of words of PP recipients.

- (15) a. Als de speaker die treffer abusievelijk aan Amokachi
 when the speaker that hit mistakenly to Amokachi
 toekent, grijpt hulptrainer Jo Bonfrère in.
 assigns intervenes assistant-coach Jo Bonfrère PART
When the speaker mistakenly assigns the goal to Amokachi, the assistant-coach intervenes.
- b. Niemand kan aan de Westduitse bondskanselier de heen-
 nobody can to the West German president the to
 en terugreis voorschrijven.
 and from-journey prescribe
Nobody can prescribe both ways of the journey to the West German chancellor.

The non-canonical word order is by far the least frequent realization of the dative construction, with less than five percent of the data falling in this category. It is nevertheless possible to identify certain factors that increase the chance of finding this alternant. We first investigate the influence of weight on the order of the NP and PP argument.

3.5.1 Weight

If we simply count the average number of words for all shifted and canonical PP recipients, we find that the canonical PP recipients are much heavier than the shifted PPs, as shown in table 3.6. In addition, the canonical direct objects are lighter. This is in line with the principle 'light precedes heavy'.

However, we need to distinguish the relative ordering in the midfield from extraposition phenomena, which are known to be influenced by syntactic weight. In V2 sentences, it is impossible to see whether the second pole, the right edge of the midfield follows or precedes the PP. Any sentence is thus ambiguous between a canonical word order sentence and an instance of PP extraposition. To ensure that the PP is in the midfield and has not been extraposed, we only include sentences in which the PP is followed by

Dataset			OBJ1	OBJ2	OBJ1/OBJ2
CGN	NP PP	(N=56)	2.02	1.93	1.05
Alpino	NP PP	(N=17)	2.53	2.29	1.10
TwNC	NP PP	(N=100)	3.27	2.17	1.51
CGN	PP NP	(N=7)	4.71	2.71	1.73
Alpino	PP NP	(N=8)	3.50	3.25	1.08
TwNC	PP NP	(N=101)	2.64	2.70	0.98

Table 3.7: Average number of words of non-extraposed PP recipients.

a verbal cluster, which indicates the right pole. Controlling for this factor, the weight effect disappears. However, the amount of data we have is very small, especially for the inverted word order. This is caused by the effect of combining various constraints on the data. PP recipients are less frequent than NP recipients and the shifted word order is less frequent than the canonical word order. If we combine these constraints and furthermore restrict ourselves to unambiguously non-extraposed PPs, we are left with very little data. This data sparseness can be overcome with automatically annotated data. We extracted all shifted PP patterns from the automatically annotated TwNC corpus and checked the results manually. This resulted in a total of 101 examples. We furthermore extracted the first 100 correct examples of the unshifted construction and compared the average weight of direct and indirect object in these test sets. The results in table 3.7 show that direct object and indirect object are almost equally heavy. Comparing the shifted and the unshifted construction, we see that the direct objects are heavier in the shifted construction than in the canonical argument order, and the indirect objects are lighter in the shifted construction than in the canonical argument order. This is contrary to what is expected, based on the complexity principle.

The large difference between the weight of unambiguously non-extracted PP recipients and those that may have been extracted shows that the influence of weight on argument ordering is not a linearization constraint: it did not influence the ordering of the two objects in the middle field and it does not influence the relative order of the NP and the PP in the middle field in dative PP constructions. Instead, it is a preference for heavy constituents not to be embedded in the VP, but to be extraposed instead.

Dataset	OBJ1			OBJ2		
	Indef	Def	Prons	Indef	Def	Prons
TwNC NP PP (N=100)	44%	56%	6	22%	78%	5
TwNC PP NP (N=101)	85%	15%	0	36%	64%	0

Table 3.8: Pronouns, definites and indefinites in the dative PP construction.

3.5.2 Pronominality and definiteness

That leaves open the question of what influences the relative order of the NP and the PP argument. The obvious candidates: information structure and pronominality. The first has often been proposed in the literature as a factor influencing word order (Gundel, 1988; Prince, 1992; Arnold et al., 2000) and the second (related) factor was already shown to be the crucial factor determining the relative order of the objects in the middle field of the double object construction. Table 3.8 list the numbers of pronouns, definites and indefinites in the shifted and unshifted (manually checked) datasets.

The table shows that the percentage of indefinite direct objects is much higher in the shifted OBL OBJ1 construction (85%) compared to the canonical OBJ1 OBL construction (44%), in line with the principle that says definite precedes indefinite. However, it would be incorrect to conclude that we have a constraint DEF<INDEF or NEW (Choi, 1996) that outranks the constraint on canonical wordorder. It is true that in a little over over half the non-canonical examples (55%), the direct object was indefinite and the indirect object was definite, in which case the non-canonical word order could be accounted for by the definiteness principle. But 32% of the canonical examples also had indefinite direct objects and definite indirect objects. And since the canonical construction is many times more frequent than the shifted construction, this means that the majority of the examples with an indefinite OBJ1 and a definite OBJ2 examples is in the canonical argument order. Instead we do assume a constraint NEW, but it is generally ranked below CANON. The distance between the two constraints must be small, though, so that there is some probability of NEW outranking CANON under a stochastic implementation of OT. This constraint then also explains why we find no direct object pronouns in the non-canonical construction: even if NEW outranks CANON, it simply does not apply to pronouns, as they are by definition old information.¹⁰

(16) NEW: [-NEW] should precede [+NEW] Choi (1996)

¹⁰It does not, however, account for the lack of pronominal *indirect* objects. We leave this issue for future research.

Among the non-canonical examples, we find an interesting subgroup. A number of examples consists of expressions containing semantically light verbs which form a collocation with their direct object: *aandacht geven/schenken*, *excuses aanbieden*, *gehoor geven*, *uitdrukking geven*, *leiding geven*, ‘give attention’, ‘offer excuses’, ‘give attention’, ‘give expression’, ‘give leadership’ (17) and this tendency of the direct object to appear at the right pole of the midfield appears to increase with the strength of the collocation. We can illustrate this by looking at two examples which mean almost the same, *aandacht geven/schenken* and *gehoor geven* ‘give attention’. *Gehoor* is in the lexicon as an independent noun, but it is most often used with the verb *geven* ‘to give’ and thus forms a very strong collocation with *geven*. This is not true for *aandacht*, which forms a weak collocation with the verbs *geven* and *schenken* ‘to give’. The stronger collocation appeared twice in canonical order in the complete 75M word corpus vs. 6 times in the shifted construction, the weaker collocation appeared 40 times in the canonical word order vs. 17 times shifted. This is as predicted by the Inherence Principle (Haeseryn et al., 1997) and the idea of semantic connectedness (Wasow, 2002) in section 3.2: the closer the verb and the direct object are connected, the higher the chance of finding the non-canonical order in which the direct object immediately precedes the verb.

- (17) Het is zeer uitzonderlijk dat het Israëlische leger in een dergelijk
 it is very exceptional that the Israeli army in a such
 geval aan nabestaanden zijn excuses aanbiedt.
 case to relatives his excuses offers
 It is very exceptional for the Israeli army to offer an apology to the
 relatives in such a case.

To conclude, the dative PP construction is most frequently realized in the canonical word order. Two factors were shown to increase the chance of finding the non-canonical order: definiteness and light verb constructions. However, it is not generally the case that in those restricted contexts, the non-canonical variant is more frequent than the canonical one. Expressed in OT terms, we must conclude that the constraint on canonical word order is the highest ranked relevant constraint, but certain lower ranked constraints are close enough to outrank canonical word order with some (low) frequency. There remain some non-canonical examples in which neither of these factors is present. We leave the question as to why we find the inverted order in these cases for further research.

In the last two sections, we looked at the order of the arguments in the double object construction and in the recipient PP construction. It was

shown that pronominality determines the order in the double object construction and that definiteness and “inherence” influence the argument order in the dative PP construction. The next section deals with the question what determines the choice for an NP or a PP recipient.

3.6 The NP/PP Alternation

The third and last alternation in the Dutch dative construction concerns the choice for an NP or PP recipient argument. As the syntactic category does not have to influence the linearization of arguments in Dutch, we do not expect to find influence from general linearization principles here. Instead, we expect to find constraints that directly influence the realization of the recipient. One such constraint is the selectional restriction of verb classes, as proposed in Levin (1993).

3.6.1 Lexical preferences

Levin (1993) argues that some English verb classes select for one variant in the dative alternation and some verbs select for the other. As the argument order variation and the NP/PP alternation go together in English, these selectional restrictions influence both argument order and the syntactic function of the recipient. In Dutch, we would expect that such construction specific constraints only influence the realization of the recipient argument, not the order of the arguments, which is assumed to be governed by more general linearization constraints.

We tested the influence on verb class on the argument variation in the double object construction on the one hand and the NP/PP alternation on the other hand by calculating the association between verb lexeme and word order/recipient type in the annotated Alpino and CGN corpora. We express this association with a log-likelihood score (see section 1.2.3 for a description of the log-likelihood measure) and regarded each verb as a separate class. As we already saw that the argument order is influenced by the syntactic category of the direct object, we controlled for this by calculating the influence separately for *het*-objects, pronominal objects and NP objects. Table 3.9 summarizes the log-likelihood scores for both alternations in all three object classes. We see that the association between verb lexeme and word order is not significant in any of the OBJ1 classes. The association between verb lexeme and the NP/PP alternation does reach significance in two classes out of three classes, suggesting that the verb lexeme does influence the choice for an NP or PP recipient. These results tell us that the distribution of NP and

Alternation	Obj1	Degrees of Freedom	LL	Significant
Arg Order (NP NP)	NP	35	6.2	no
	pron	20	22.9	no
	het	7	4.4	no
NP/PP Alternation	NP	40	79.8	p=0.001
	pron	24	36.5	p=0.050
	het	7	8.3	no

Table 3.9: Loglinear (LL) scores for the relation between verb lexeme and surface form in different OBJ1 categories.

Alternation	Ent before	Ent Cat		Ent Verb		Ent both	
NPNP order	0.172	0.110	-36%	0.152	-12%	0.094	-45%
NP/PP recipient	0.578	0.578	-0%	0.426	-26%	0.422	-27%

Table 3.10: Entropy of the word order and dative NP/PP alternation

PP recipients over the different verbs cannot be attributed to chance. This does not necessarily mean that it is due to the different lexical preferences.

To get a better idea of the relative impact of direct object category on the one hand and the verb lexeme on the other hand on the argument order and the NP/PP alternation, we calculated the entropy of the system, based on the automatically parsed TwNC corpus. The results are in figure 3.10. This entropy is a measure of the uncertainty about whether or not shift will apply (or whether to realize the recipient as an NP or a PP). We first calculated the entropy of the system without adding any information. This starting entropy is low, as the canonical word order is much more frequent than the shifted word order. We then added either information about the syntactic category of the direct object (*het*, personal pronoun or NP) or the verb lexeme. The entropy reduction after adding OBJ1 information is much higher than after adding verb lexeme information, even though there are many more verb lexeme categories than OBJ1 categories. We did the same calculations for the NP/PP alternation. Here, we see the reverse picture: adding OBJ1 category information does not reduce the entropy of the system, but adding verb lexeme information leads to an important entropy reduction. Adding information about both the verb lexeme and the direct object category leads to an entropy reduction as big as the sum of the entropy reductions of the two pieces of information individually. This indicates that the OBJ1 category and the verb lexeme are independent variables.

Adding verb class information reduced the entropy considerably, confirming our hypothesis that the construction specific lexical preferences only influence the NP/PP alternation, not argument ordering. The entropy did not go down to zero, but it is not expected to, as some verbs allow free variation. Nevertheless, it is not excluded that other factors influence the choice for one NP recipients over PP recipients or vice versa. We check the influence of two more general constraints: weight and pronominality.

3.6.2 Weight and pronominality

If we look at the average weight of indirect objects in table 3.5 on page 85, we see that the recipient arguments that are realized in PPs are much heavier than those that are realized as NPs. But these numbers may not be measuring an effect of weight on the NP/PP alternation proper: they may be a side effect of the constraints on DOS identified in section 3.4. We saw that even in Dutch, word order and the NP/PP alternation are not completely independent. In the double object construction, the theme-recipient order is available for pronominal themes only. Thus, in sentences with heavy recipients and full NP themes, realization as a PP is the only way in which the recipient can be “moved” to the right.

In order to look at the NP/PP alternation proper, we have to control for this and ensure that the order of the arguments in the sentence does not play a role. We decided to restrict ourselves to PP recipients which co-occur with an inanimate pronominal direct object. In these sentences, the direct object may precede both an NP and a PP indirect object in the midfield. We thus controlled for influences of ordering effects. Furthermore, we only included PP recipients which undeniably precede the second pole (i.e. precede the verb cluster). This excludes extraposition effects.

For the annotated data (Alpino and CGN), that left us with only 13 examples. To collect a more representative data collection, we extracted from the automatically parsed TwNC corpus 200 sentences that met our criteria, and manually checked them for parse errors. For NP recipients, we constructed a similar dataset. For the annotated data, we took the CGN data, as 12 of the 13 PP examples were from CGN. In addition, we again extracted 200 examples from the automatically annotated corpus, again restricting ourselves to double object constructions with pronominal objects. The table in 3.11 still shows a striking difference in weight between NP and PP recipients

Although we know now that the difference in weight must be associated with the difference in syntactic category, not with the order of the arguments it still does not prove that there is an influence of the Complexity Principle on the NP/PP alternation. The distributional differences in the double object

Dataset	NP		PP	
Alpino+CGN	1.12	N=52	1.53	N=13
TwNC	1.43	N=200	2.35	N=200

Table 3.11: Average weight for PP and NP recipients in the midfield

construction appeared to be weight-driven, but turned out to be based on pronominality. The same applies to the dative PP construction. There were 20 pronouns among the PP recipients vs. 161 among the NP recipients. If we remove pronominal recipients from the data (leaving 180 sentences in the PP category and only 39 in the NP category) we find that the PP recipients are in fact lighter (2.5 words) than the NP recipients (3.1 words). Weighing of the non-pronominal NP recipients on a larger set of 100 instances reduced the average to 2.7 words, which is however still heavier than the PP recipients. The conclusion must be that there is not so much a difference in weight between NP and PP recipients, but rather a pronominality difference. We did one more test to illustrate the effect of pronominality on the NP/PP alternation: we compared the percentage of one word recipients taken up by pronouns in the automatically parsed corpus (thus controlling for weight effects). We find 22% (20041/92553) pronouns in the one word NP recipients vs. 6% (347/5527) in the PP data. In the annotated CGN data this contrast is less pronounced, but still clearly there: 82% (581/706) vs. 49% (52/106). All together, this shows that pronominal recipients disprefer realization as an oblique. This effect cannot be reduced to the general linearization constraints we saw in section 3.4 (PRON<NP): in our data, all direct objects were pronominal, so that the order with the recipient following the theme was possible both with an NP and a PP recipient. Many of our examples were in fact shifted. In other words: in this restricted domain, realization as an NP or a PP does not determine the relative order of the arguments.

Finally, recall from the literature section that Bresnan and Nikitina (2003) argued that local person NPs should be realized as objects, not obliques. As a local person NP is always realized as a pronoun, this could explain why we find many more pronominal NP recipients than pronominal PP recipients. But no evidence for such a restriction was found: of the 20 pronouns in our 200 sentence PP test set, 12 (60%) were local. Of the 161 pronouns in the NP test set, 84 (52%) were local.¹¹ We also tested the existence of

¹¹Note that third person inanimate PP recipients were excluded. These are realized as R-pronouns, which we filtered out (see section 3.3.1). This could potentially influence the data, especially since Bresnan et al. (2005) argue that inanimate recipients prefer the PP structure. However, as the vast majority of recipients is animate (in fact, *all* NP

	local	3rd person
NP NP _{unshifted}	101	52
NP NP _{shifted}	13	9
NP PP	7	3
PP NP	2	1

Table 3.12: Person features of pronominal indirect objects

a person effect in our annotated data (without any further restrictions on direct object category or extraction). Table 3.12 shows the distribution of third person and local pronouns over the four alternants. The differences between the frequencies of third and local person pronouns is not significant ($p=0.05$). Even after aggregating together the counts for the two orderings in the double object construction and the two orders in the dative PP construction, we still did not reach significance. Although the numbers are too small for definite conclusions, they do suggest that person does not influence the NP/PP alternation.

3.6.3 Implementation in OT

We have identified two factors that influence the dative NP/PP alternation in Dutch: lexical preferences and pronominality. In addition, we saw in earlier that the dative PP construction is less frequent, over all. In this section, we incorporate these findings in our OT model. First, we look at Bresnan and Nikitina’s account of the English alternation (Bresnan and Nikitina, 2003). They adopt the set of constraints in (18).

(18) Constraints on the Dative Alternation (Bresnan and Nikitina, 2003)

*STRUCT: avoid syntactic structure (here: *PP)

FAITH(REC): express the recipient role of a verb with distinct marking (case or adposition)

OO-PRIMACY: OBJ2 strictly dominates OBJ1 on hierarchies of informational prominence.

recipients were animate), we do not expect that ignoring this factor has a large influence on the person effect.

HARMONY(1,2): *NP_{Noun} & *PP_{1,2Person}.

*STRUCT can be applied to both English and Dutch to account for the skewed overall distribution of NP and PP recipients; we simply adopt this constraint. We have found no evidence of an influence of OO-PRIMACY on the NP/PP alternation in Dutch. However, pronominality and information structure did play a role in the linearization of arguments. It would be interesting to see how many of the effects that have been analyzed as OO-PRIMACY effects could be accounted for with our linearization constraint, given the strict argument ordering in English. However, if the constraint proves indispensable, this is fully compatible with our account for Dutch (which would have the constraint ranked below all relevant constraints).

This is not the case for HARMONY(1,2). Bresnan and Nikitina's harmony constraint penalizes pronominal recipients realized as a dative PP, which is what we want for Dutch, too. However, it also penalizes realizations for which we found no evidence of dispreference in Dutch, such as full NP OBJ2s. We propose therefore to have the simplex constraint *PP_{pro} instead.¹² Note that our account excludes the existence of the harmony constraint. By definition, constraint conjunctions outrank both their conjuncts, but our data contradict a constraint HARMONY(1,2) which outranks *PP: such a constraint would predict a dispreference for full NP indirect objects, which we did not find. Our account thus clashes with the account of Bresnan and Nikitina under the assumption of a universal set of constraints. It is an interesting question whether the general linearization constraints we identified can be used to account for the English data without the use of HARMONY(1,2). We leave this for future research.

That leaves the question of how to implement the lexical preferences. Bresnan and Nikitina assume that FAITH(REC) is parameterized for different classes of ditransitive verbs. These parameterized constraints are then ranked at various positions in the hierarchy, as in the constraint ranking in (19) (incorporating only a few of the lexical preferences), so that different frequencies are predicted for the PP and NP realizations of their recipient arguments.

- (19) Constraint Ranking for the Dative Alternation (Bresnan and Nikitina, 2003)

¹²This constraint could be derived from aligning the pronominality hierarchy with the Core and Noncore argument hierarchy in the same way the two conjuncts of the HARMONY(1,2) constraint were derived, using the techniques familiar from the work of Aissen (1999, 2003).

$$\begin{aligned} &\text{OO-PRIMACY} \gg \text{FAITH}(\text{Rec})_{yell} \gg \text{HARMONY}(1,2) \\ &\gg \text{FAITH}(\text{Rec})_{fax}, \text{FAITH}(\text{Rec})_{give} \gg * \text{STRUCT} \end{aligned}$$

We adopt this approach in order to illustrate how lexical preferences interact with the other constraints. However, some remarks should be made about this approach. The parameterized faithfulness constraints are unlikely to be universal. As such, they are in violation of the basic assumption in OT that the set of constraints is universal and that the only source of variation is the ranking of these constraints (Prince and Smolensky, 1993). At the same time, Smolensky and Legendre (2005) acknowledge that this principle may have to be weakened to account for certain language particularities. An alternative is to introduce a lexical feature which specifies the strength of the recipient. The constraint could then refer to this feature instead of the verb (class) itself. Such an approach crucially relies on a lexicon friendly OT model (van der Beek and Bouma, 2004). The two approaches make different predictions: in the first model, all members of a (semantically motivated) class have the same distribution, whereas the latter allows variation within such classes. There appear to be such differences, but at the same time, Lapata (1999) showed that Levin's verb classes do have empirical value. This regularity would be unaccounted for in the second model.

In tableau 3.13, we illustrate how lexical preferences interact with the other constraints. We included only one constraint $\text{FAITH}(\text{REC})$, which favors the PP realization. We rank it on a par with $*\text{PP}$, so that free variation is expected if no other constraint penalizes one of the constructions. An example of such a verb is *betalen* 'pay'. We expect a 50-50 division between NP and PP recipients if both arguments are NPs. If we have a pronominal recipient, $*\text{PP-PRO}$ prevents it from being realized as a PP resulting in a double object construction.

- (20)
- a. Bo betaalt de student tien euro.
Bo pays the student ten euros
 - b. *Bo betaalt tien euro de student.
Bo pays ten euros the student
 - c. Bo betaalt tien euro aan de student.
Bo pays ten euros to the student
 - d. ??Bo betaalt aan de student tien euro.
Bo pays to the student ten euros
 - e. ??Bo betaalt de student het.
Bo pays the student it

- f. ??Bo betaalt het de student.
Bo pays it the student
- g. Bo betaalt het aan de student.
Bo pays it to the student
- h. Bo betaalt aan de student het.
*Bo pays to the student it
- i. Bo betaalt hem tien euro.
Bo pays him ten euros
- j. *Bo betaalt tien euro hem.
Bo pays ten euros him
- k. ??Bo betaalt tien euro aan hem.
Bo pays ten euros to him
- l. ??Bo betaalt hem tien euro.
Bo pays to him ten euros
- m. ??Be betaalt hem het.
Bo pays him it
- n. Bo betaalt het hem.
Bo pays it him
- o. ??Bo betaalt het aan hem.
Bo pays it to him
- p. *Bo betaalt aan hem het.
Bo pays to him it

At least two other groups of verbs exist. Verbs like *verwijten* ‘blame’ prefer the double object construction, unless there is some external reason for realizing the recipient as a PP, for example to avoid ambiguity (see section 3.7). For these verbs, the faithfulness constraint either does not apply or it is ranked below the markedness constraint *PP. Finally, there are verbs such as *verkopen* ‘sell’, which generally prefer to realize the recipient argument as a PP, unless it is a recipient. Here, the appropriate FAITH(REC) must outrank the markedness constraint *PP. The minimal constraint ranking accounting for these three classes is thus as in (21).

(21) Minimal Constraint Ranking for the Dutch Dative Alterations

$$\begin{aligned}
 &(\text{PRO}_{it} < \text{NP/PRO}) \gg (\text{PRO}_{3rd} < \text{NP}) \gg *PP\text{-PRO} \gg \\
 &\quad \text{FAITH(Rec)}_{verkopen} \gg *PP, \text{FAITH(Rec)}_{betalen} \gg \\
 &\quad (\text{OBJ2} < \text{OBJ1} < \text{OBL}) \gg \text{NEW}
 \end{aligned}$$

Input: <i>pays</i> (<SUBJ><OBJ1><OBJ2>)		PRO _{it} <NP / PRO	PRO _{3rd} <NP	*PP-PRO	*PP FAITH(Rec)	OBJ2 < OBJ1 < OBL	
OBJ1='ten euros' OBJ2='the student'	☞ NP NP _{unshift}				*		(20-a)
	☞ NP NP _{shift}				*	*!	(20-b)
	☞ NP PP				*		(20-c)
	☞ PP NP				*	*!	(20-d)
OBJ1='it' OBJ2='the student'	☞ NP NP _{unshift}	*!			*		(20-e)
	☞ NP NP _{shift}				*	*!	(20-f)
	☞ NP PP				*		(20-g)
	☞ PP NP				*	*!	(20-h)
OBJ1='ten euros' OBJ2='him'	☞ NP NP _{unshift}				*		(20-i)
	☞ NP NP _{shift}				*	*!	(20-j)
	☞ NP PP			*!	*		(20-k)
	☞ PP NP			*!	*	*	(20-l)
OBJ1='it' OBJ2='him'	☞ NP NP _{unshift}	*!			*		(20-m)
	☞ NP NP _{shift}				*	*	(20-n)
	☞ NP PP			*!	*		(20-o)
	☞ PP NP			*!	*	*	(20-p)

Table 3.13: Competition between the four alternants of the Dutch dative alternation.

3.7 More Factors in the Dative Construction?

We have focused on three factors that influence the dative alternations: weight, pronominality/definiteness and lexical preferences. No doubt many more factors influence the realization of ditransitive verbs in one way or another.

Bresnan et al. (2005) argue that animacy is a relevant feature. Unfortunately, none of the available corpora of Dutch is annotated with information about animacy, making it impossible to test this hypothesis on corpus data. Within the restricted search space of the pronominal recipients, there were too few inanimate recipients to draw any conclusions. That being said, it does seem to be the case that with (marked) inanimate recipients, the DOS is less acceptable and the PP-construction is preferred. This is illustrated in the constructed examples in (22).

- (22) a. Ik geef dit boek een tien.
 I give this book a ten
 I give this book ten out of ten.
 b. ?Ik geeft dat geen enkel boek
 I give that no single book
 I do not give that to any book.
 c. En toch geef ik dat wel aan dit boek.
 and still give I that indeed to this book
 But I still do give that to this book.

The markedness of (22-b) may also be explained by the aim to avoid ambiguity. If both objects are inanimate, it is harder to tell which one is the direct object and which one the indirect. In such cases, the canonical word order or dative marking with a PP appear obligatory. Similarly, if the direct object is an atypical object, e.g. a local (1st or 2nd person) pronoun, it is easily misunderstood as an indirect object (which is often local). By marking the real indirect object with the preposition *aan*, this reading is excluded and ambiguity is reduced.¹³ This would explain the ungrammaticality of the constructed example in (23-a) and the grammaticality of (23-b).¹⁴

¹³This could also be regarded a mild OO-PRIMACY effect.

¹⁴Note that the shifted version (i) is also ungrammatical under the intended reading. This was already accounted for in section 3.4. Even if it were grammatical, it would not solve the ambiguity.

- (i) %als ik jou Ajax verkoop.
 if I you Ajax sell

- (23) a. *als ik Ajax jou verkoop.
 if I Ajax you sell
 b. als ik jou aan Ajax verkoop.
 if I you to Ajax sell
 If I sell you to Ajax.

There are indications that the surface string also has some influence on the realization of the recipient argument. Among the sentences with PP recipients, for example, we find many that have proper name recipients, proper name agents and non-pronominal themes. As DOS is only available for pronouns, only the canonical double object construction is possible. This canonical argument order would lead to two proper names in a row (24-a). Realizing the recipient as a PP argument successfully avoids this sequence of proper names (24-b).

- (24) a. Daar gaf volgens de overlevering God Mozes het
 there gave following the tradition God Moses the
 gebod "Gij zult niet stelen".
 commandment thou shalt not steal
 b. Daar gaf volgens de overlevering God aan Mozes het
 there gave following the tradition God to Moses the
 gebod "Gij zult niet stelen".
 commandment thou shalt not steal
 *Tradition has it that this is the place where God gave Moses the
 commandment "Thou shalt not steal".*

Finally, we argued in section 3.2 that the account of Reinhart (1996) in terms of stress and focus is based on dubious data and cannot account for all data. At the same time, we noted that there are reasons to expect an influence of focus on linearization. This influence is not restricted to the dative alternation and should be studied with a corpus that is annotated with information structure. Unfortunately, such a corpus is not available for Dutch.

In the previous sections, pronominality and definiteness constraints were shown to override the canonical word order in the Dutch dative alternation in some instances and the NP/PP alternation was shown subject to lexical preferences and—surprisingly—pronominality constraints. This section discussed some additional influences which appear to influence the dative alternation. The most important constraints were modeled in the OT framework. The next section provides some additional evidence for an important part of this model, namely the ordering constraints on nominals. This evid-

ence comes from a second construction which is sensitive to the pronoun scale: the Accusativus cum Infinitivo.

3.8 Additional Evidence: the AcI

In section 3.4 it was shown that a pronoun hierarchy exist in Dutch, similar to the one for German (Müller, 2001). It was furthermore shown that DOS is sensitive to this scale: the weaker the pronoun, the more prone it is to shift. This was formalized in a set of linearization constraints. In this section we will digress from the dative alternation to illustrate that the various constraints for aligning pronouns of different strengths can also be applied to other word order alternations. We show how these constraints account for the distribution of embedded object shift (EOS) in the Accusativus cum Infinitivo (AcI) construction.

The AcI construction illustrated in examples (25) and (26) in and figure 3.1 is headed by a sensory verb, the verb *laten* (to let) or the verb *helpen* (to help). The verb takes an object and an xCOMP. The embedded subject is functionally controlled by the object.

- (25) Op haar elfde zag ze Russische tanks haar land
 On her eleventh saw she Russian tanks her country
 binnenvallen.
 invade
At age eleven she saw Russian tanks invade her country.

Several LFG analyses of this construction exist, e.g. Bresnan et al. (1982), Zaenen and Kaplan (1995) and Kaplan and Zaenen (2003). All nominal arguments (also the embedded ones) are selected for in the VP, all verbal arguments in V', thus accounting for the crossing dependencies that occur when one AcI constructions is embedded in another, as illustrated in a well known example from the literature (26):

- (26) omdat ik Cecilia Henk de nijlpaarden zag helpen voeren.
 because I Cecilia Henk the hippos saw help feed
because I saw Cecilia help Henk feed the hippos.
- (27) C-structure rules for the AcI-construction Kaplan and Zaenen (2003)
- $$VP \rightarrow \begin{matrix} NP^* \\ (\uparrow \text{xCOMP}^* \text{ OBJ}) = \downarrow \end{matrix} V'$$

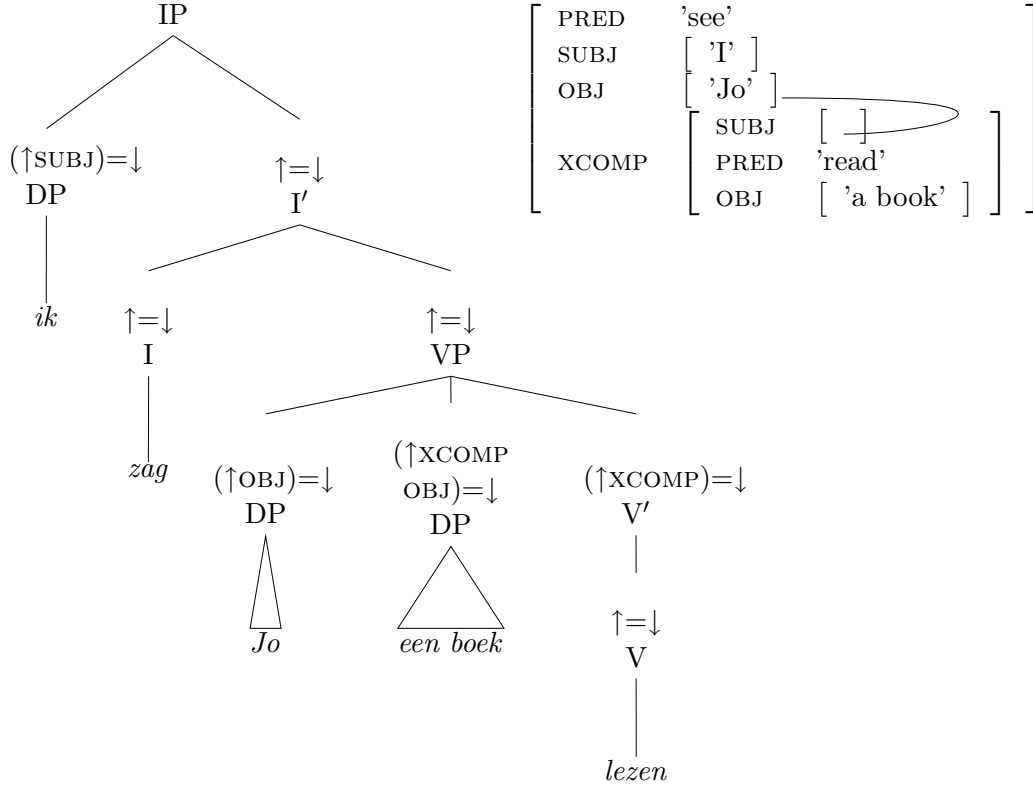


Figure 3.1: C-structure and f-structure for an AcI-construction in Dutch

$$V' \rightarrow V \left(\begin{array}{c} V' \\ (\uparrow \text{XCOMP}) = \downarrow \\ (\uparrow \text{XCOMP}^+ \text{OBJ}) \not\prec_f (\uparrow \text{OBJ}) \end{array} \right)$$

The order of the nominal arguments is restricted to the canonical word order in (25) and (26) by the f-precedence requirement $(\uparrow \text{XCOMP}^+ \text{OBJ}) \not\prec_f (\uparrow \text{OBJ})$ (Kaplan and Zaenen, 2003). This constraint states that the constituent that maps onto the embedded OBJ1 in the f-structure cannot precede the constituent that maps onto the f-structure of the main clause direct object. However, under certain conditions, the embedded object can shift over the higher object (or embedded subject) (28). In other words: the f-precedence constraint is violable. The conditions under which we find EOS resemble the conditions on DOS. A difference is that DOS was only blocked with local person pronouns, while EOS is blocked with all animate pronouns. This is best illustrated with animate and inanimate examples of the weak pronoun *ze* (them) (29-b)-(29-a). Note that inanimate objects are very unmarked. More marked objects have to stay in their canonical object position.

- (28) a. Ik zag 't Jo doen.
 I saw it Jo do
 I saw Jo doing it.
- b. Ik zag 't 'r doen
 I saw it her do
 I saw her doing it.
- c. Ik zag dat haar ouders doen.
 I saw that her parents do
 I saw her parents doing that.
- d. Ik zag haar ouders dat doen.
 I saw her parents that do
 I saw her parents doing that.
- e. Ik zag ze dat doen.
 I saw them that do
 I saw them doing that.
- f. Ik zag het ze doen.
 I saw it them do
 I saw them doing it.
- (29) a. Ik heb ze Jo door zien slikken.
 I have them Jo through seen swallow
 I saw Jo swallowing them.
- b. %I heb ze Jo zien zoenen.
 I have them Jo seen kiss
 I saw Jo kissing them.

The constraints on EOS resemble those on DOS. Again we see an interaction between pronouns that want to precede full NPs and canonical word order preventing this. In contrast to DOS, we now have the dividing line between animate and inanimate pronouns. This is modeled by ranking the canonical word order constraint for the AcI construction between the linearization constraints for animate and inanimate pronouns. The canonical word order constraint (30) states that complement functions (here restricted to OBJ1, OBJ2 and OBL) precede complement functions that are embedded in in an XCOMP.

- (30) CF < XCF: OBJ1, OBJ2 and OBL precede (XCOMP OBJ1), (XCOMP OBJ2) and (XCOMP OBL)

The rest of the analysis works as for the OS in the double object construction, as illustrated in table 3.14: if both the OBJ1 and (XCOMP OBJ1) are full

Input: <i>see</i> (<SUBJ><OBJ1><XCOMP>)			PRO _{it} <NP / PRO	PRO _{3rd/inanim} <NP	CF <XCF	PRO _{3rd/anim} <NP	O2 < O1 < OBL	PRO _{loc} < NP
OBJ1='Jo'		OBJ1 XOBJ1						
XOBJ1='a book' ex.(25)	✎	XOBJ1 OBJ1			*!			
OBJ1='Jo'		OBJ1 XOBJ1	*!					
XOBJ1='it' ex.(28-a)	✎	XOBJ1 OBJ1			*			
OBJ1='her parents'		OBJ1 XOBJ1		*!				
XOBJ1='that' ex.(28-c)	✎	XOBJ1 OBJ1			*			
OBJ1='them'		OBJ1 XOBJ1						
XOBJ1='that' ex.(28-e)	✎	XOBJ1 OBJ1			*!			
OBJ1='them'		OBJ1 XOBJ1	*!					
XOBJ1='it' ex.(28-f)	✎	XOBJ1 OBJ1			*			
OBJ1='Jo'		OBJ1 XOBJ1				*		
XOBJ1='them' ex.(29-b)	✎	XOBJ1 OBJ1			*!			

Table 3.14: Embedded Object Shift in the AcI

NPs, we simply get the canonical word order, as none of the linearization constraints on pronouns fires. The same happens if both arguments are pronouns, but not local or *het*. However, if the direct object is *het* or it is another pronoun and the indirect object is a full NP, the linearization constraints on pronouns, which outrank the constraint on canonical word order, determine that we get an EOS.

Now what happens if we embed a ditransitive in an AcI construction? In principle both types of shift are possible: within the XCOMP argument, direct objects may shift in front of the indirect objects (DOS), and in addition, the embedded object may shift in front of the direct object of the main verb (EOS). Some of the possibilities are listed in (31). For three examples, we put the constraint interaction in tableau 3.15.

- (31) a. Ik heb de docent de leerlingen het boek zien geven.
 I have the teacher the students the book see give
I saw the teacher giving the students the book.

Input: <i>see</i> (<SUBJ><OBJ1><XCOMP>)		PRO _{it} <NP/Pro	PRO _{3rd/inanim} <NP	CF<XCF	PRO _{3rd/anim} <NP	O2<O1<OBL	PRO _{loc} <NP
OBJ1=NP XOBJ1=NP XOBJ2=NP (31-a)	OBJ1 XOBJ2 XOBJ1 OBJ1 XOBJ1 XOBJ2 XOBJ2 OBJ1 XOBJ1 XOBJ2 XOBJ1 OBJ1 XOBJ1 XOBJ2 OBJ1 XOBJ1 OBJ1 XOBJ2			*! *!* *!* *!		*! * *	
OBJ1=NP XOBJ1=het XOBJ2=NP (31-b)	OBJ1 XOBJ2 XOBJ1 OBJ1 XOBJ1 XOBJ2 XOBJ2 OBJ1 XOBJ1 XOBJ2 XOBJ1 OBJ1 XOBJ1 XOBJ2 OBJ1 XOBJ1 OBJ1 XOBJ2	*! *! *! *!		* ** **! *		* * *	
OBJ1=NP XOBJ1=NP XOBJ2=pron (31-c)	OBJ1 XOBJ2 XOBJ1 OBJ1 XOBJ1 XOBJ2 XOBJ2 OBJ1 XOBJ1 XOBJ2 XOBJ1 OBJ1 XOBJ1 XOBJ2 OBJ1 XOBJ1 OBJ1 XOBJ2			*! *!* *!* *!	* * * *	*! * *	

Table 3.15: Tableaux for (31)

- b. Ik heb het de docent de leerlingen zien geven.
I have it the teacher the students see give
I saw the teacher giving it to the students.
- c. Ik heb de docent hun het boek zien geven.
I have the teacher them the book see give
I saw the teacher giving them the book
- d. %Ik heb hun de docent het boek zien geven.
I have them the teacher the book see give

The model predicts that it is possible to shift an inanimate (XCOMP OBJ2) in front of the direct object of the main verb. Example (32) shows that this

results in bad sentences, contrary to what we expect. The ungrammaticality of (32) may be explained by the fact that inanimate indirect objects are rare. One may expect a ‘worst of the worst effect’ (Smolensky, 1995; Lee, 2003), resulting in the ungrammaticality of the combination of both a marked indirect object and a marked argument order.¹⁵

- (32) a. Ik zie hem het boek een tien geven
 I see him the book an A give
 I see him give the book an A.
 b. ??Ik zie het hem een tien geven
 I see him that an A give
 I see him give it an A.

3.9 Conclusion

We investigated the influence of various factors on the dative alternation in Dutch, all of which have been claimed to influence the dative alternation in English. These factors are weight, definiteness, pronominality and lexical preferences. The first three factors are considered general linearization principles. These linearization principles may influence the dative alternation in English, as the order of the arguments alternates with the syntactic category of the recipient. This is not the case in Dutch. The Dutch data thus allowed us to study the argument order and the NP/PP alternation in isolation.

We expected to find a split in the factors influencing the different aspects of the dative alternation in Dutch: linearization constraints influencing the argument order alternations, and construction specific constraints influencing the NP/PP alternation. This expectation is partially borne out. Lexical preferences of the verb indeed only influenced the choice for an NP or a PP recipient, not the order of the arguments. And pronominality and definiteness were indeed shown to influence the order of the arguments.

But not all expectations were borne out. Pronominality was assumed to be a linearization constraint, related to the definiteness constraint. It was thus predicted to have an influence on argument order but not on the NP/PP alternation. Corpus data however showed that pronominality does influence the NP/PP alternation (as well as argument order). This is in line with work based on harmonic alignment of the nominal hierarchy and the semantic role hierarchy (Silverstein, 1976; Aissen, 1999; Bresnan et al., 2005, for example).

¹⁵In an OT model, such effects may be modeled by means of a constraint conjunction. In this case, a conjunction of a (low ranked) constraint penalizing inanimate objects and a constraint penalizing non-canonical word order. The conjunction of two constraint always outranks both component constraints.

Syntactic weight is another classic linearization constraint. But surprisingly, syntactic weight did not influence the order of the arguments in the midfield, neither in the double object construction nor in the dative PP construction. It was shown that extra weight does increase the chance on finding extraposition.

Pronominality and definiteness did influence the argument order in the dative alternation, as expected. Pronominality and pronoun type determined whether DOS applied and definiteness was shown to have a mild influence on the relative order of the direct object and the dative PP. In general, there was a strong preference for the canonical argument orders, independent of the syntactic category of the recipient.

Formalizing the constraints on argument order in an OT setting allowed us to illustrate the ranking and interaction of the constraints. We thus showed how the most frequent patterns could be predicted. But the dative alternation shows a lot of variation, with different realizations occurring with extremely skewed distributions. Classic OT cannot account for these patterns. We sketched how a stochastic interpretation of OT, STOT Bowersma and Hayes (2001), could account for those less frequent realizations of the dative construction. It would be interesting to see whether the frequencies predicted by an actual implementation in STOT would match the frequency distributions that we observed in the corpora.

The NP/PP alternation was shown to be subject to lexical preferences. It remains an open question how these preferences are best modeled in OT. Lexical variation appears to be a serious problem for the general assumption in OT that all variation is driven by the ranking of universal constraints.

Chapter 4

Determinerless PPs

In this chapter we will use a large, automatically parsed corpus to extract the lexical information needed to facilitate an account of determinerless PPs in a (computational) grammar of Dutch. Determinerless PPs (PP-Ds) are a heterogeneous group of constructions which pose problems for formal and computational grammars. We will describe the different types of PP-Ds and indicate how a grammar could account for them. For these accounts, information is required about the prepositions and nouns that participate in the construction. This information is not generally available, but with the use of corpus data a base-repository of PP-Ds is generated semi-automatically.

4.1 Introduction

Combinations of prepositions and singular nouns show many of the problematic characteristics of multiword expressions (MWEs): the syntax and semantics of the construction is often—but not always—idiosyncratic, and at the same time the constructions are to some degree productive or allow modification (Baldwin et al., 2003). With these characteristics, PP-Ds pose problems for formal and computational grammars: the grammar should allow *op reis* ‘on journey’ but not **op stoel* ‘on chair’ or **ik maak reis* ‘I make journey’. It should analyze *in zwang* ‘in fashion’, but not allow the string *zwang* in any other context. It should not parse or generate the unmodified **op wijze* ‘on way’, even if the modified *op slinkse wijze* ‘on sneaky way’ is fine and other PP-Ds allow both modified and unmodified versions (*op (lange) termijn*) ‘later/long-term’, lit. ‘on (long) term’.

In this chapter we present a general characterization of PP-Ds and we describe ways in which different types of the construction could be handled by a grammar. We will see that all of these accounts share the prerequisite

that information about which prepositions and which nouns participate in which type of PP-D, as well as the modifiability of the P-N combination, should be available. Unfortunately, this is generally not the case. The second half of this chapter will show how this lack of information can be overcome by using a large, automatically parsed corpus to compose the lexical resource semi-automatically. Section 4.2 is partly based on Baldwin et al. (2003) and Baldwin et al. (to appear).

4.2 The Syntax of Determinerless PPs

In earlier work (Baldwin et al., 2003) it was shown that PP-Ds do not form a homogeneous group. They argued that in principle, each combination of a preposition and a singular noun without a determiner is a PP-D, but these P-N combinations differ with respect to their syntactic and semantic markedness. In addition, it may be either the noun or the preposition which selects for the lack of a determiner. We will argue that at least one more distinction has to be made, namely whether or not the PP as a whole is dependent on a verbal or nominal head.

A PP-D minimally consists of a preposition and a noun. If this noun is a plural or an uncountable noun (*suiker* ‘sugar’), which in itself may constitute a saturated noun phrase, the resulting structure is syntactically unmarked (*met suiker* ‘with sugar’). But if the noun is a countable noun, which in itself does not constitute a saturated noun phrase (**huis is mooi* ‘house is beautiful’), the resulting structure (*naar huis* ‘to house’) is syntactically marked. In this chapter the focus will be on syntactically marked determinerless PPs and from now on we will use the term PP-D to refer to this subset of all P-N combinations. Although there are criteria for distinguishing countable from uncountable nouns, e.g. only countable nouns co-occur with numerals, and only uncountable singular nouns combine with *veel* ‘much’, distinguishing countable from uncountable nouns is not a simple task. More about countability information and distinguishing marked from unmarked determinerless PPs in section 4.3.2 and more about countability classification in general in chapter 5.

The absence of a determiner may come with idiosyncratic semantics, such as in *buiten spel* ‘not in a position to influence the matter’, lit. ‘offside’. Although we will mention semantic effects in PP-Ds on occasion, the focus of this chapter is on the syntactic properties of PP-Ds. For a more thorough discussion of the semantics of PP-Ds, we refer to Stvan (1998) and Baldwin et al. (to appear).

The syntactically marked PP-Ds may be further subdivided in four classes:

the fully fixed PP-Ds, PPs with bare noun NPs, compositional PP-Ds and prepositions selecting for determinerless NPs. The following four sections each discuss the syntactic properties of one class of PP-Ds. Furthermore, each type of PP-D may be either independent or dependent on a verbal (or a nominal) head.

4.2.1 Fixed determinerless PPs

The first type of PP-D is the class of fully fixed PP-Ds such as *in zwang* ‘in fashion’. Modification of this type is either excluded or fully fixed, as in *naar *(eigen) zeggen* ‘according to him/herself’, lit. ‘after own saying’, and the construction is non-compositional and non-productive. This class contains PP-Ds which contain lexical items that do not occur outside of PP-Ds (anymore). For example, the word *a* from the PP-D *a priori* is not used as a preposition in regular, compositional PPs. Similarly, *lieverlee*, from the construction *van lieverlee* ‘gradually’ is not used as a noun elsewhere. The strings *a* and *lieverlee* can be classified as a preposition and a noun on the basis of information from other languages or historical variants of Dutch, but do not behave as such in present-day Dutch.

Lexical listing is a simple and sufficient solution for this type of PP-D. Instead of breaking the string down into a preposition and a noun and including both separately in the lexicon—which would incorrectly predict both items to occur without the other—we analyze the string as a word with spaces. The syntactic category may be adjective or adverb (1), depending on the syntactic contexts in which the PP-D occurs, and additional annotations may be added where appropriate. For example, we added the annotation (\uparrow ATYPE) = pred in (1) to indicate that the PP-D is only used in predicative constructions. The string as a whole is associated with a unique predicate, which does not bear any formal relation to the meaning of the subparts.

- | | | | | |
|-----|------------------------|-----|---------------------|-------------------|
| (1) | <i>in zwang</i> : | A | (\uparrow PRED) | = ‘in_zwang’ |
| | | | (\uparrow ATYPE) | = pred |
| | <i>van lieverlee</i> : | Adv | (\uparrow PRED) | = ‘van_lieverlee’ |

In order to include these words with spaces in the lexicon, one needs a repository of PP-Ds which are fully fixed. Section 4.3 is devoted to the semi-automatic construction of such a repository.

4.2.2 Independent bare noun NPs

A second type of PP-Ds consists of a preposition and a bare noun NP. An example of such a bare noun NP is *school* in the English *in/at/after school*.

Stvan (1998) describes different types of what she calls defective NPs in English, and focuses on the semantic effects of the determinerless use in PPs. For example, the ‘institutionalized location denoting nouns’ such as *school* in (2-a) do not refer to the building as such, but to the related activity, in this case attending classes. This is illustrated by the fact that it is not appropriate to use the determinerless construction for the mayor visiting the school. The crucial difference between these PP-Ds and others is that the noun occurs without a determiner outside of PPs as well, and in these sentences we observe the same semantic effects as Stvan observed for the PP-Ds (2-b) (Baldwin et al., 2003). That means that strictly speaking, the construction is not syntactically marked, but comparable to a PP with a bare plural object.

- (2) a. John is at school.
b. School is over.

PPs with bare noun NPs are much more common in English than they are in Dutch (or German). Furthermore, there is little evidence for a particular semantics associated with this type of PP-D in Dutch. In the category of the institutionalized location denoting nouns we find *school* ‘school’ and *kantoor* ‘office’, which can both be used determinerless to refer to an activity, but only one of them (*school*) can be used determinerless outside of PPs: *kantoor* is only used without a determiner in PP-Ds (3) and is thus not a bare noun NP. It is better analyzed as part of a compositional PP-D, a type that we will discuss in the next section. For other classes of bare noun NPs, for example crime names in Dutch, no particular semantic effect of the absence of the determiner is observed: there is no difference in meaning between *doodslag* ‘homicide’ in example (4-a) and in example (4-b).

- (3) a. Meteen na kantoor wandel ik met de hond.
directly after office walk I with the dog
I walk the dog directly after work.
b. *Ik vind kantoor niet leuk
I like office not PART
I do not like work
- (4) a. Hij is veroordeeld wegens doodslag.
he is convicted for homicide
He has been convicted for homicide.
b. De rechter acht doodslag niet bewezen.
the judge holds homicide not proved
The judge holds homicide not proved.

To account for PPs with bare noun NPs, no special machinery is needed. The defective noun phrases may be special, but the combination of this noun phrase and a preposition is the same as for other uncountable words. We assume a grammar for Dutch which is equivalent to the toy grammar in (6) with respect to PP-Ds. We will use the sample lexical entries in (5), which capture those aspects of Dutch nouns and determiners that are relevant in the PP-D analysis, while abstracting away from the characteristics that are not of direct interest here.

We assume an NP analysis for all nominals. The determiner contributes the attributes DETFORM and DETTYPE to the NP. The value of the first is the surface string of the determiner. The latter has one of the values ‘definite’, ‘indefinite’, ‘demonstrative’ or ‘null’. The value ‘null’ is explained in detail below, the other values are for indefinite, definite and demonstrative determiners respectively. By setting this value, the existential constraint on the noun in the NP rule is satisfied. This constraint states that the f-structure projected from the noun (i.e. the NP) should be defined for DETTYPE.

Parentheses around a functional equation indicate that it is optional.¹ For example, the determiner is optional in the NP rule. This facilitates NPs without a determiner, for example NPs with uncountable or plural nouns. But these NPs still need to be defined for DETTYPE to satisfy the existential constraint on the noun. Countable nouns cannot satisfy this constraint, because they do not have this feature. As a result, they cannot constitute an NP by themselves. But uncountable (mass) nouns and plurals are optionally defined DETTYPE=indef, either in the lexicon or via a lexical/morphological rule, and thus satisfy the DETTYPE constraint on NPs, even if no determiner is present.

(5)	<i>auto</i> :	N	(↑PRED)	= 'auto'
	<i>suiker</i> :	N	(↑PRED)	= 'sugar'
			((↑DETTYPE)	= indef)
	<i>een</i> :	D	(↑DETFORM)	= een
			(↑DETTYPE)	= indef
	<i>de</i> :	D	(↑DETFORM)	= de
			(↑DETTYPE)	= def

¹The optionality can be resolved by writing two separate entries, one with the optional equation and one without.

$$\begin{array}{lcl}
 (6) & NP & \Rightarrow \begin{array}{cc} (D) & N \\ \uparrow=\downarrow & \uparrow=\downarrow \\ & \uparrow_{\text{DETTYPE}} \end{array} \\
 & PP & \Rightarrow \begin{array}{cc} P & NP \\ \uparrow=\downarrow & \uparrow=\downarrow \end{array}
 \end{array}$$

Words that can form independent bare noun NPs, such as *school* in Dutch and English and crime names like *moord* ‘murder’ in Dutch are assigned a lexical entry similar to uncountable nouns (7), allowing for occurrence with and without a determiner in and outside of PPs. The optionality of the DETTYPE annotation ensures that definite NPs such as *de school* ‘the school’ are still allowed.

$$\begin{array}{lcl}
 (7) & school: & (\uparrow_{\text{PRED}}) = \text{'school'} \\
 & & ((\uparrow_{\text{DETTYPE}}) = \text{indef})
 \end{array}$$

4.2.3 Compositional determinerless PPs

The third type of PP-D is syntactically marked: it consists of a preposition (e.g. *in*) and a true count noun, which only occurs without a determiner inside a PP, such as *termijn* ‘term’. The meaning of the sentence is composed from the regular meaning of the preposition and the regular meaning of the noun, but in some cases semantic effects occur. For example, the meaning of *naar huis* means ‘to one’s own house’, ‘home’. The PP must be headed by a particular preposition (8-a) or a member of a particular set of prepositions (8-b). The productivity of this construction varies from nouns occurring without a determiner with only one preposition to nouns that occur with a wide range of prepositions, but productivity is never unrestricted.

- (8) a. op/*in/*na termijn
on/in/after term
- b. in/op/naar/*onder/*naast bed
in/op/to/under/next to bed

Modification may be excluded (9-a), optional (9-b) or obligatory (9-c). Orthogonal to this three-way distinction, the modification may be restricted (9-d) or virtually unrestricted (9-e), resulting in 5 different modification patterns.

- (9) a. in *zacht bed
in soft bed

- b. op (hoog) niveau
on high level
- c. op *(slinkse) wijze
on sneaky way
- d. op ski-/*lange/*mooie vakantie
on ski/long/beautiful vacation
- e. op vegetarische/politieke/water-/\dots basis
on vegetarian/political/water-/\dots basis

The possibility for these nouns to occur in PP-Ds, and the restrictions that apply to the determinerless occurrences of these nouns, can be represented in their lexical entries. First, we allow for an optional DETTYPE annotation, which will satisfy the f-structure constraint in the NP rule. The value of DETTYPE is ‘null’. In this, the structure differs from the syntactically unmarked PPs consisting of a preposition and a bare plural, which are annotated DETTYPE=indef. The null value facilitates an implementation of semantic effects such as a familiarity effect (Stvan, 1998), which contradict a value ‘indef’: *naar huis* ‘to house’ does not mean to some indefinite house, but to one specific house, namely one’s own (i.e. home). These effects disappear if a determiner is added.

Secondly, we conjoin this optional DETTYPE annotation with restrictions on the syntactic category of the mother (PP), the head of the phrase, and the presence of adjuncts. If the optional annotation is instantiated, then all conjoined annotations are instantiated as well.

If modification is not allowed, e.g. for *bed* (9-a), the noun cannot have a feature ADJ. This is indicated by the negated existential constraint in (10). The lexical entry for *bed* further requires that the PP it is contained in is headed by one of the prepositions *in* ‘in’, *op* ‘op’ or *naar* ‘to’. This is achieved by the inside-out function application $((\text{OBJ}\uparrow) \text{PTYPE}) = \{\text{in}|\text{op}|\text{naar}\}$, that states that the f-structure of which ‘bed’ is the OBJ should have a PTYPE ‘in’, ‘op’ or ‘naar’.

Termijn ‘term’ (9-b) can only form a PP-D with the preposition *op* ‘on’. The lexical entry is optionally specified DETTYPE=null, and if this is an annotation is instantiated, the PTYPE of the PP has to be ‘op’. The noun can optionally be modified by any modifier, so no further constraints are necessary. Note, though, that the heavier the modification, the less acceptable the PP-D becomes, and the higher the probability of finding a determiner. We do not account for this heaviness effect here.²

²This heaviness effect may be accounted for with a general Optimality Theoretic constraint. Alternatively, one could specify the possible type(s) of modification on each lexical entry separately. For instance, one could restrict modification to attributive adjectives only

Wijze ‘way’ is obligatorily modified, but the modifier is unrestricted (9-c). This is modeled by the existential constraint (\uparrow ADJ). Again, the heaviness constraint applies here, effectively ruling out postnominal modification.

(10)	<i>bed</i> :	N	(\uparrow PRED)	= ‘bed’	
			{((OBJ \uparrow) PTYPE)	= _c {in op naar}	&
			(\uparrow DETTYPE)	= null	&
			$\neg(\uparrow$ ADJ)}		
	<i>termijn</i> :	N	(\uparrow PRED)	= ‘term’	
			{((OBJ \uparrow) PTYPE)	= _c op	&
			(\uparrow DETTYPE)	= null}	
	<i>wijze</i> :	N	(\uparrow PRED)	= ‘way’	
			{((OBJ \uparrow) PTYPE)	= _c {in op naar}	&
			(\uparrow DETTYPE)	= null	&
			(\uparrow ADJ)}		

More complicated are the compositional PP-Ds with restricted modification. The noun *evenwicht* ‘balance’ can form a PP-D with the preposition *in* ‘in’. There is only one adjective that can (optionally) modify the noun: *wankel* ‘unstable’. Similarly, only the adjective *onmiddellijke* ‘immediate’ can modify the noun *ingang* ‘start’, but now the modification is obligatory (*met* **(onmiddellijke) ingang* ‘immediately’, lit. ‘with immediate start’).³ In the lexical entry of the noun, we constrain the predicate of its adjunct as in (11). In case the modification is optional, we only state that the modifier cannot have a PRED other than the fixed modifier *wankel* ‘unstable’. For obligatory modification, we again restrict the predicate value of the adjunct and additionally state that the NP must have a feature ADJ. The only way to satisfy both constraints is to realize the fixed modifier, in this case *onmiddellijke* ‘immediate’.⁴

with a constraint (ATYPE = ‘attributive’) on the adjunct.

³*Ingang* ‘start’ can also be followed by a PP headed by *van* ‘from’ (*met ingang van* ‘starting’, lit. ‘with start from’). This is assumed to be a collocational preposition, but could alternatively be analyzed as a PP modifier in the PP-D, in which case the obligatory modifier is either the adjective with PRED ‘immediate’ or a PP with PTYPE ‘from’.

⁴This formalization still allows for multiple occurrences of the lexically selected modifiers. It is difficult to get definite grammaticality judgments for those sentences, but it may well be that they have to be considered out. We believe that this is not a characteristic of PP-Ds, but rather a more general phenomenon that penalizes literal repetitions of words. Tracy H. King (p.c.) furthermore noted that it is technically possible to avoid having more than one adjunct by blocking the presence of the attribute SCOPE, which determines the relative scoping of the adjuncts.

- (11) *evenwicht*: N (\uparrow PRED) = 'balance'
 $\{((\text{OBJ}\uparrow) \text{ PTYPE}) =_c \text{ in} \quad \&$
 $(\uparrow \text{ DETTYPE}) = \text{null} \quad \&$
 $\neg((\uparrow \text{ ADJ} \ni \text{ PRED}) \neq \text{'unstable'})$
ingang: N (\uparrow PRED) = 'start'
 $\{((\text{OBJ}\uparrow) \text{ PTYPE}) =_c \text{ met} \quad \&$
 $(\uparrow \text{ DETTYPE}) = \text{null} \quad \&$
 $(\uparrow \text{ ADJ}) \quad \&$
 $\neg((\uparrow \text{ ADJ} \ni \text{ PRED}) \neq \text{'immediate'})$

In our account of compositional PP-Ds, non-heads (nouns, NPs) pose restrictions on their head (the preposition). A similar approach was advanced by Soehn and Sailer (2003) in HPSG. This work focuses on so called *unique nominal complements*, nouns which never occur outside of PPs. This is in contrast to our case of compositional PP-Ds, where the nouns do occur independently as well as inside PPs. Ideally, one would generate and analyze the use of the noun in and outside of PPs as instances of one and the same lexical entry, avoiding duplication of identical information (e.g. PRED information) in the lexicon. The LFG framework facilitates such an analysis via the mechanism of optional (complexes of) f-structure annotations: if the noun is used outside of a PP, the optional annotations are not realized. As a consequence, it is not defined for DETTYPE and can only form an NP after combining with a determiner. If the noun occurs inside a PP-D, the optional annotations have to be instantiated to satisfy the DETTYPE constraint, and consequently the structure has to satisfy all other conjuncts of the complex of optional annotations, restricting the head of the PP and modification. We see again that in order to implement an account of this type of PP-Ds, we need detailed information about which nouns combine with which prepositions in a PP-D and the modifiability of the resulting construction.

4.2.4 Prepositions selecting for determinerless NPs

In the fourth and last type of PP-D, it is the preposition that licenses determinerless NP complements. This type of PP-D again has two subtypes. On the one hand we have the prepositions *per* 'per', *ter* and *ten* 'at' (with different archaic case-markings), which obligatorily select for a PP-D (12).⁵

⁵There is one use of *per* with a determiner:

- (i) Ik stop per de eerste van de volgende maand.
 I quit from the first of the next month
I will quit on the first of next month.

In the case of *per*, this is a fully productive process. *Ter* and *ten* are historically contractions of the preposition *te* ‘at’ and an article, and their use is restricted. *Te* is still used productively, but only with city names, which are always determinerless. Some fixed expressions with *te* do contain determiners, e.g. *te allen tijde* ‘at all times’ (but *te voet* ‘on foot’ and *te midden van* ‘amidst’). This is in contrast with *ten* and *ter*, which occur in many fixed combinations, but never with a determiner.

On the other hand we have prepositions that license PP-Ds, but can occur with regular NPs as well, such as *zonder* ‘without’ (13). Sometimes, the determinerless occurrences are restricted to a certain semantic domain. An illustration of this phenomenon is the preposition *in* ‘in’, which combines with any bare noun indicating a piece of clothing, but requires a ‘regular’ NP elsewhere (14).

- (12) a. Ik zal het per koerier laten bezorgen
 I will by courier let deliver
 I will have a courier deliver it.
 b. *Ik zal het per een koerier laten bezorgen
 I will by a courier let deliver
- (13) a. Ik kan best zonder auto.
 I can fine without car
 I can live just fine without a car.
 b. Ik kan best zonder een auto.
 I can fine without a car
 I can live just fine without a car.
- (14) a. Zie je die student in
 see you that student in
 pak/uniform/spijkerbroek/bloemetjesjurk?
 costume/uniform/jeans_{sg}/flower dress?
 Do you see that student in costume/uniform/jeans/flower dress?
 b. *Ken je die student in kerk/gebouw/klas/groep/kamer?
 Know you that student in church/building/class/group/room
 c. *Ik vind pak/uniform/spijkerbroek/bloemetjesjurk niet
 I consider costume/uniform/jeans_{sg}/flower dress not
 mooi.
 nice

Prepositions selecting for ‘real’ PP-Ds should be distinguished from preposi-

We do not account for this use of *per* in this chapter.

tions that often occur with uncountable nouns. An example of such a preposition is *wegens* ‘on account of’ or the collocational preposition *op verdenking van* ‘on suspicion of’, which occur very frequently with the names of crimes, which we saw can form NPs by themselves 4.2.2. If one combines these prepositions with true count nouns, the determiner is again obligatory (15).

- (15) a. A. werd veroordeeld wegens een autokraak
 A. was sentenced one account of a car break-in
 A. was sentenced on account of breaking into a car
 b. *A. werd veroordeeld wegens autokraak
 A. was sentenced on account of car break-in

Treating *per* and *zonder* ‘without’ as regular prepositions, the PP-Ds in (12-a) and (13-a) violate the grammar rules in (6): the bare count nouns are not specified for DETTYPE and thus cannot form an NP. But prepositions do not combine with bare nouns or N’s. We solve this by specifying the value for the nouns DETTYPE attribute *on the preposition*. By doing so, the count nouns can form an NP, but only if this NP functions as the complement of this particular preposition.⁶

The preposition *per* specifies its complement to be DETTYPE=null (16-a). Thus NPs with a determiner, as well as bare plurals, mass nouns and proper names, are correctly excluded. For *zonder* ‘without’, this would not be correct: example (13-a) and (13-b) are both grammatical and can be used interchangeably. We model this with an optional annotation ((↑OBJ DETTYPE)=indef). All this annotation does is providing the necessary DETTYPE attribute if it is not provided by a determiner. The value is the same as for NP complements with an indefinite determiner, correctly predicting identical semantics for example (13-a) and (13-b). The optionality of the annotation ensures that definite prepositional complements can still be derived.

- (16) a. *per*: P (↑PRED) = ‘per(OBJ1)’
 (↑OBJ DETTYPE) = null
 b. *zonder*: P (↑PRED) = ‘zonder(OBJ1)’
 ((↑OBJ DETTYPE) = indef)

A different approach is necessary for semantically restricted PP-Ds. One could try to add the semantic restriction on the lexical entry of the preposition, but this would imply that *all* members of the semantic class allow for the determinerless construction. This appears not to be the case. Compare (14-a) with (17-a) and (17-b) with (17-c).

⁶This solution is supported by the fact that the prepositions *ter* and *ten* are historically a combination of a preposition and an article.

- (17) a. ??Zie je die student in broek?
 see you that student in trousers_{sg}
 b. Hij is op reis/tournee/pad/weg/vakantie/expeditie.
 he is on voyage/tour/path/way/vacation/expedition
 c. *Hij is op trip/tocht.
 he is on trip/journey

Alternatively, one can treat each example as an instance of a compositional PP-D. Although this appears to be empirically correct, it does not capture the semantic generalization. We will get back to these semantically restricted PP-Ds in section 4.3.5.

4.2.5 Determinerless PPs as dependents

Orthogonal to the distinctions made in Baldwin et al. (2003, to appear), a number of (Dutch) PP-Ds function only as dependents of a verbal (or nominal) head. In this case, the preposition does not combine with determinerless nominals unless it is an argument of this particular verbal or nominal head. The PP-D itself may be fully fixed (18-a) or compositional (18-b). In some cases, the head of the prepositional complement selects for any nominal object, as long as it is determinerless (18-c)-(18-d). We will refer to dependent PP-Ds as determinerless prepositional complements. Dependent PP-Ds should be distinguished from independent PP-Ds, because their distribution is much more restricted. The lexical entry which licenses the occurrence of such PP-D is not the entry of the preposition or the noun in this PP-D, but the entry of the verb (or the noun) which selects for the PP-D complement: they will be analyzed as fixed prepositional arguments Villada Moirón (2004). Table 4.1 presents an overview of the different types of PP-Ds that we have distinguished.

- (18) a. Iemand in toom houden
 someone in bridle hold
 To restrain someone
 b. In première gaan
 in premier go
 To have its opening night
 c. Van auto veranderen
 of car change
 Change cars
 d. De functie van voorzitter
 the function of president

The function of president

4.3 Extraction of PP-Ds

4.3.1 Introduction

In the previous section we identified various types of PP-Ds: fully fixed and compositional PP-Ds, bare noun NPs and bare noun selecting prepositions, all of which could be independent PPs or dependents of verbal/nominal heads. We indicated how each of these types could be accounted for in a grammar. Although the analyses differ in many respects, they all share one prerequisite: that information is available about which nouns and which prepositions participate in which type of PP-D and which kind of modification is allowed.

This information is not available, generally. The Dutch part of Euro-WordNet Vossen and Bloksma (1998) lists a total of seven PP-Ds. The electronic dictionaries Celex (Baayen et al., 1993) and Parole⁷ do not include any information on PP-Ds. Even the Dutch reference grammar (Haeseryn et al., 1997) and the main dictionary (Geerts and Heestermans, 1992) do not include information about PP-Ds systematically, although the grammar includes a list of collocational prepositions (i.e. fixed combinations of a preposition, a—possibly determinerless—NP and another preposition). Only the Alpino lexicon, which is part of the Alpino parser (Bouma et al., 2001; van der Beek et al., 2002b) contains more information about marked PPs. The lexicon lists about 95 fixed PP-Ds, such as *a priori* ‘a priori’, *in feite* ‘in fact’ and *van nature* ‘by nature’ and 132 (semi-automatically extracted) collocational prepositions consisting of a preposition, a bare noun and another preposition, such as *in antwoord op* ‘in reply to’, (Bouma and Villada, 2002). Furthermore, almost fifty fixed parts of larger idiomatic expressions, for example phrasal verbs, contain PP-Ds.

The work presented in this section aims at overcoming the lack of systematic lexical information by means of (semi)automatic extraction of PP-Ds from corpus data. We want to identify prepositions that select for determinerless NPs, nouns that participate in compositional PP-Ds, and fixed P-N tuples. Furthermore, subcategorized PP-Ds should be distinguished from independent ones. Like regular uncountable nouns, the independent bare noun NPs from section 4.2.2 will be discarded.

⁷<http://www.inl.nl/corp/parole.htm>

Type	Example	Characteristics
Selection by P (obligatorily)	<i>per te vroeg geboren kind</i> (per premature child)	modification allowed high prep-noun entropy no PP+D
Selection by P (optionally)	<i>zonder hoed</i> (without hat)	modification allowed high prep-noun entropy also PP+D
Fixed PP-D	<i>in principe</i> (in principle)	zero modification entropy high verb entropy no PP+D
Idiosyncratic P-N pairs	<i>op straat</i> (on street)	no NP-D restricted modifiability high verb entropy
Bare noun NPs	<i>wegens moord</i> (for murder)	also NP-D high preposition entropy high verb entropy
PP-D phrasal verbs	<i>in premiere gaan</i> (in premier go, have opening night)	low dep-head entropy
PP-D verbal complements	<i>van auto veranderen</i> (of car change)	low dep-head entropy

Table 4.1: Overview of PP-D types.

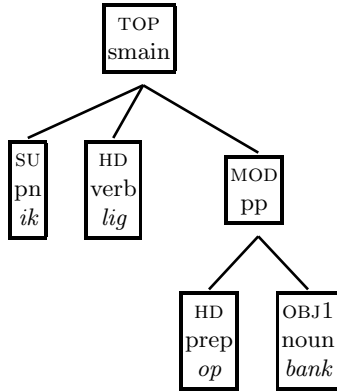
4.3.2 Preliminaries

The PP-D extraction methods proposed in this work are all based on parsed corpus data. We used a 75M word newspaper corpus that was automatically annotated with dependency structures by the Alpino parser (Bouma et al., 2001; van der Beek et al., 2002b). Examples of dependency trees are in example (19)-(21). The parser has an overall accuracy of 85.5%, measured over the dependency relations. The syntactic annotation allows us to extract information about (different types of) modification and about the verb that governs the PP-D. Both types of information are difficult to extract with shallower forms of annotation.

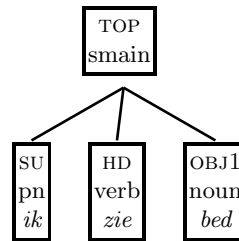
Baldwin et al. (2003) argued that chunked data should be preferred over parsed data for extraction of PP-Ds. The motivation for this is that the quality of the output of the parser is conditioned on the ability of that parser to analyze PP-Ds, giving rise to circularity. However, in our case we used the Alpino parser, for which this circularity does not arise. The Alpino parser overgenerates in that it generally allows bare nouns to form an NP. Although this is not always correct, it allows every P+N combination to be analyzed as a PP-D, even if the noun in itself does not constitute a saturated noun phrase. The only PP-Ds that will not be retrieved systematically are those that feature in collocational prepositions or function as a fixed part of a larger idiomatic expression, because these are analyzed as a single lexical item. This analysis as a word with spaces is based on the annotation guidelines for collocational prepositions of the Corpus of Spoken Dutch (Moortgat et al., 2001). As a result, these PP-Ds will not be retrieved by querying for P+N patterns. Instead, the collocational prepositions will show up in the results as prepositions.

As a result of the overgeneration strategy of the Alpino parser, implementing a detailed account of PP-Ds will not improve coverage: Alpino already assigns the grammatical *op reis* ‘on journey’ a PP analysis. However, it assigns the same parse to ungrammatical PP-Ds (19) and to the ungrammatical use of the bare noun outside of the PP-D (20). If one aims at a parser which parses all and only grammatical strings or a generator which generates all and only grammatical sentences, one needs to replace the overgenerating NP \Rightarrow N rule by a detailed account of PP-Ds (and an extensive list of uncountable nouns, a named entity recognition module, etc.).

- (19) *Ik lig op bank.
I lie on couch

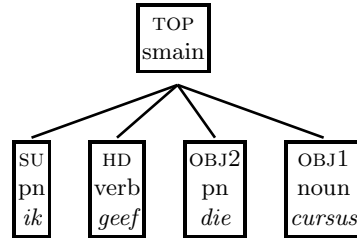


- (20) *Ik zie bed.
I see bed



Such a restriction to the grammar is expected to reduce ambiguity and improve accuracy of the parser, as incorrect application of the NP \Rightarrow N rule may result in incorrect full parses of grammatical sentences. For illustration, if regular count nouns are prohibited to form a NP by themselves, then the noun *cursus* ‘course’ cannot form an NP and the incorrect parse in (21) is ruled out.

- (21) Ik geef die cursus
I give that course
I teach that course



From the parsed data we extracted the preposition heading the PP-D, the noun object of the preposition and the verb that heads the PP-D.⁸ The nouns were restricted to singular common nouns. Determiners and appositions were not allowed in the NP. Furthermore, we extracted the list of modifiers (both post- and prenominal) and the list of dependency triples corresponding to the heads of the modifiers (allowing for generalizations over types of modifiers).

As the goal of this research is the extraction of syntactically marked PP-Ds, we are only interested in PP-Ds with a countable noun, which in itself cannot constitute an NP. Unfortunately, the availability of reliable count-

⁸Or the verb that heads the NP in which the PP-D is embedded. We included these verbs in order to recognize fixed verbal arguments that are misanalyzed as NP internal modifiers.

ability information is limited. Nouns in the Alpino lexicon (14K words) are labeled with countability information. We found a 81.1% agreement between the countability judgments in Alpino and a 196 word hand annotated gold standard. The judgments in the gold standard were based on actual occurrences of the nouns in the Twente Nieuws Corpus.⁹ In addition to this, we used a list of 6K nouns that were automatically classified as countable, uncountable or both according to the method described in Baldwin and van der Beek (2003) and chapter 5 of this thesis. The accuracy of this list is 85.7%. Note furthermore that nouns may have both countable and uncountable usages. In this research, any noun which has at least one uncountable sense is considered syntactically unmarked whenever it occurs without a determiner (in or outside of a PP). For example, *buiten beeld* ‘offscreen’ is not considered syntactically marked, because *beeld* has an uncountable use e.g. *goed beeld hebben* ‘to have good reception’. The reason for this approach is that word sense disambiguation is not yet feasible in broad scale computational grammars, thus countable and uncountable senses cannot be distinguished from each other.

Using the extracted data, we first identify prepositions that (optionally) select for determinerless objects, then fully fixed PP-Ds and nouns that select for occurrence in compositional PP-Ds. Finally, we illustrate how the same methods can be applied to extract a set of PP-Ds that are selected for by particular verbs, forming phrasal verbs or determinerless prepositional complements.

4.3.3 Prepositions selecting for determinerless NPs

It was shown in section 4.2.4 that the prepositions that select for determinerless objects can be subdivided in prepositions that only occur in PP-Ds and those that optionally select for a determinerless NP. The prepositions *ter* and *ten* ‘at’ (with archaic casemarkings) and *per* ‘per’ obligatorily select for PP-Ds and can simply be listed in the lexicon with the lexical entry given in example (16-a). We excluded them from our data in further experiments.

Other prepositions optionally select for a determinerless complement. To extract these prepositions, we calculated the ratio between the number of noun types in PPs with a determiner headed by preposition P and the number of types in PP-Ds headed by the same preposition. We excluded nouns that were flagged as uncountable and used a frequency cutoff of 50. This relatively high cutoff was used to filter out many of the infrequent collocational prepositions, while keeping the very frequent simplex prepositions. As

⁹<http://wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html>

Preposition	Example Noun	Ratio	Countability
<i>vol</i> ‘full of’	<i>liefde</i> ‘love’	0.00	uncount
<i>bij wijze van</i> ‘by means of’	<i>geintje</i> ‘joke’	0.01	count
<i>qua</i> ‘wrt’	<i>lichaam</i> ‘body’	0.02	count
<i>op verdenking van</i> ‘on suspicion of’	<i>moord</i> ‘murder’	0.29	uncount
<i>richting</i> ‘towards’	<i>schatkist</i> , ‘treasury’	0.37	count
<i>op het gebied van</i> ‘concerning’	<i>kunst</i> ‘art’	0.81	uncount
<i>zonder</i> ‘without’	<i>paspoort</i> ‘passport’	0.86	count
<i>als</i> ‘as’	<i>banneling</i> ‘exile’	0.90	count
<i>bij gebrek aan</i> ‘lacking’	<i>ervaring</i> ‘experience’	1.04	uncount
<i>tot</i> ‘to’	<i>bedelaar</i> ‘beggar’	1.93	count

Table 4.2: Prepositions that select for PP-Ds

the list of uncountable nouns is far from complete, many uncountable nouns still show up in the data. We therefore also extracted a sample of 15 noun objects of these prepositions and manually classified them as countable or uncountable, making it possible to distinguish prepositions that select for ‘real’ PP-Ds from those that simply select for uncountable objects. Table 4.2 lists the ten prepositions with the lowest type-ratio, including both prepositions that frequently co-occur with uncountable nouns and prepositions that allow uncountable nouns to form a determinerless prepositional object.

Finally, there are prepositions which generally select for regular NP objects, but form PP-Ds with bare singular nouns from a particular semantic class to be their object. An example is the preposition *in* with clothing items such as *pak*, *bloemetjesjurk*, *uniform*, *overhemd* ‘suit, flower dress, uniform, shirt’. We argued that semantically restricted selection by the preposition would overgenerate in section 4.2.4. We therefore treat this type of PP-D as a compositional PP-D in section 4.3.5, where we hope to capture the semantic generalization by clustering individual PP-Ds on the basis of EuroWordNet synsets.

4.3.4 Fixed determinerless PPs

The second group of PP-Ds to be extracted is the fully fixed PP-Ds. A very simple heuristic was applied: we extracted all preposition-noun tuples ($F > 10$) with ‘nouns’ that only occur grammatically in PP-Ds. At least 99% of all occurrences of the noun had to be determinerless and at least 90%

Candidate	F	Modifier	P	Fixed
in afwachting (in anticipation)	610	van (of)	0.91	✓
volgens/naar zeggen (according to)	564	eigen (own)	1.00	✓
in opspraak (compromised)	366			✓
op voorhand (in advance)	347			✓
bij uitstek (pre-eminently)	345	in (in)	0.02	✓
op jaarbasis (on a yearly basis)	264	op (on)	0.02	✓
van nature (by nature)	200			✓
in diskrediet (into disrepute)	172	bij (by)	0.01	
op straffe (under penalty)	147	van (of)	0.94	✓
naar hartelust (to ones heart's content)	115	met (with)	0.03	✓
in zwang (in fashion)	83	in (in)	0.04	✓
sinds mensenheugenis (within living memory)	77	in (in)	0.03	✓

Table 4.3: Fixed PP-Ds

of the occurrences inside a PP¹⁰. The difference between the two cutoffs is motivated by the fact that very little parse errors are made with respect to the recognition of the determiner and grouping Det + N in an NP. On the other hand, the parser does not always identify the PP correctly, so we slightly relaxed the condition on determinerless occurrence outside of PPs. Finally, we set the maximum noun-preposition entropy at 1.00 as a filter for uncountable nouns.

In total, 45 candidate fixed PP-Ds were extracted. The candidates were manually checked by three native speakers, including the author. For 36 (80%) of them, at least two of the informants indicated that they knew the noun and that it only occurred in a PP context. Nine candidates were considered false positives. Some of the true positives could be considered collocational prepositions (see section 4.3.1), but as they conform to the definition of fully fixed PP-Ds that we formulated, we included them in the results.

The modifiers of the candidate PP-Ds were extracted and the entropy of the modifier given that there is one) was calculated. Table 4.3 lists the most frequent fixed PP-Ds with their frequency, the most frequent modifier and the probability of this modifier. We included all types of modification, but abstracted away from the objects in PP modifiers: all PPs headed by a particular preposition are regarded the same. The distribution of the modifiers

¹⁰A cutoff of 3 was used for PP-Ds with a frequency lower than 30, allowing for a maximum of 3 typos or parse errors

confirms that modification is fully fixed for this type of PP-D: the probabilities of the most frequent modifier are either very high, between .90 and 1.00, or else less than .05.

4.3.5 Compositional determinerless PPs

The largest set of PP-Ds is the group of compositional PP-Ds. Clearly, the simple heuristic applied for the fixed PP-Ds will not work for PP-Ds with ‘regular’ nouns: the nouns occur outside of PPs and in regular PPs as well as in PP-Ds. Instead, we used the data from our automatically parsed corpus to calculate the association between the absence of a determiner and the occurrence in a PP. The count noun may occur without a determiner outside of PPs, for example as a result of the universal grinder or a parse error, but it will appear determinerless in PPs much more frequent than expected based on the ratio of NP/PP contexts. The association is measured with the log-likelihood ratio, implementation based on the NSP package Banerjee and Pedersen (2003).

We excluded the nouns of the fixed PP-D category, the prepositions that obligatorily select for PP-Ds and prepositions from table 4.2 that optionally select for PP-Ds. We used a frequency cutoff of 10. We further restricted our candidates by setting a verb entropy minimum of 2.00. This is aimed at excluding phrasal verbs.¹¹ Table 4.4 lists the nouns for which the association is the strongest.

We see that 7 (47%) of the top 15 are nouns that do not occur without a determiner outside of PPs, making their occurrence in PP-Ds syntactically marked.¹² Only two of the nouns that were judged uncountable were listed as such, as well as one noun (*mate*, ‘measure’) that was judged countable. With large amounts of high quality countability information, precision can increase considerably.

The members of semantic classes that select for occurrence in PP-D, such as pieces of clothing (plus the preposition *in*) in Dutch, are not represented in the output. A possible explanation is that each of these nouns is infrequent, often not passing the frequency cutoff. This problem

¹¹We included nouns which occur with a low entropy with *zijn* (‘to be’), as these tend to be PP-Ds with predicative uses, not phrasal verbs.

¹²Again, nouns were marked countable if at least two out of three informants knew the word and indicated it was strictly countable. This only reflects the basic countability of the nouns. Countable words can still be used in uncountable contexts by way of coercion, the *universal grinder* being the most well known example. Despite this, we still assume that nouns have a basic countability and that this influences the possibility to occur with or without a particular determiner.

Noun	LL	P ent	P max	Count
huis (house)	2994.1	1.08	naar (to)	✓
belang (interest)	2226.4	0.18	van (of)	
beeld (view)	2161.3	0.58	in (in)	
verwachting (expectation)	2101.0	0.77	naar (to)	
straat (street)	1904.2	0.31	op (on)	✓
voorbeeld (example)	1877.9	0.52	bij (by)	✓
druk (pressure)	1636.8	0.51	onder (under)	
plaats (place)	1604.0	1.62	van (of)	
dienst (service)	1330.6	0.20	in (in)	✓
voorkeur (preference)	1251.7	0.22	bij (by)	
principe (principle)	1206.4	0.12	in (in)	✓
kracht (strength)	1058.2	1.32	met (with)	
leeftijd (age)	1050.1	0.31	op (on)	✓
totaal (total)	973.51	0.02	in (in)	✓
school (school)	930.67	1.21	op (on)	

Table 4.4: Flexible PP-Ds

can be circumvented by clustering together all members of the semantic class, thus increasing the amount of data per item. We combined the data of all hyponyms of *reis* (journey), including *tournee*, *safari*, *vakantie*, *kruistocht* (tour, safari, vacation, crusade) in EuroWordNet and we did indeed find a positive correlation between occurring inside a PP and lacking a determiner. The log likelihood of this cluster is 259.0, which is higher than that of the most frequent member of this set, *vakantie* (vacation, 195.5), but lower than that of the number two, *reis* (journey, 290.7). However, for the semantic class *kledingstuk* (piece of clothing), we found a *negative* correlation between the lack of a determiner and being the object of a preposition. For other clusters, no appropriate hyperonym could be found in EuroWordNet. An example is the set *op verzoek/bevel/aanbevelen/aanraden/initiatief* (on request/order/advise/recommendation/initiative).

Not all nouns form PP-Ds with various different prepositions: many nouns combine with only one or two prepositions in a PP-D. This is illustrated by the low preposition entropy in table 4.4. As the nouns combine with other prepositions in regular, saturated NPs, their ability to occur in preposition specific PP-Ds is not (optimally) reflected in the results of table 4.4, which generalize over all prepositions. Instead, we should measure for each noun the association between the absence of a determiner and the occurrence in a PP headed by a specific preposition. Again, the association was measured

Tuple	LL	Ent	Mod max	Count
naar huis (to house)	4972.6	0.23	met (with)	✓
van belang (of interest)	4299.2	1.86	groot (great)	
op straat (on street)	2927.6	0.30	in (in)	✓
onder druk (under pressure)	2865.3	1.28	van (of)	
naar verwachting (to expectation)	2846.2	0.21	over (about)	
in dienst (in service)	2756.5	1.18	bij (at)	✓
bij voorbeeld (for example)	2515.3	1.22	in (in)	✓
in principe (in principle)	2075.2	0.33	voor (for)	✓
bij voorkeur (by preference)	1783.7	1.01	in (in)	
op bezoek (at visit)	1744.3	0.74	bij (at)	
op leeftijd (at age)	1725.9	4.48	jong (young)	✓
in totaal (in total)	1706.3	0.27	voor (for)	✓
na afloop (after ending)	1440.5	0.33	in (in)	✓
in werkelijkheid (in reality)	1285.3	0.18	in (in)	
voor rekening (on account)	1276.6	0.91	eigen (own)	✓

Table 4.5: Flexible PP-Ds

with the log likelihood ratio. The results are listed in table 4.5.

Table 4.5 also list for each PP the modifier entropy and the most probable modifier. As expected, modification is somewhat more flexible than in the fully fixed PP-Ds. However, we see that modification is still very much restricted.

Comparing the tables 4.4 and 4.5 we see that the association calculation per preposition leads to a higher number of syntactically marked PP-Ds in the top 15: 9 (60%) instead of 7 (47%). Furthermore, looking at the 50 highest ranked nouns for both methods we see that the nouns that were found by the more general approach are almost exclusively uncountable nouns (*contrast*, *grond*, *lucht* ‘contrast, ground, air’, among others). In contrast, the nouns that were retrieved by the preposition specific method but not by the general approach are almost all count nouns (*gesprek*, *reis*, *slot* ‘conversation, journey, lock’). We conclude that the preposition specific method is less sensitive to the availability of high quality countability information, although many uncountable nouns were still retrieved.

With the preposition specific approach, performance also increases with respect to EuroWordNet clustering: the combination of *op* (‘on’) with the journey-cluster now scores 947.1, whereas *reis* (journey) and *vakantie* (vacation), the two highest ranked members of the cluster, score 748.7 and 549.3. For the clothing class, the effect is not so strong: although the negative association is no longer present, we also do not find a strong positive association

PP-N	LL	Verb ent	Verb max
tot hand (to hand)	135.7	0.00	ga (go)
buiten functie (of duty)	142.3	0.00	stel (put)
bij stuk (at place)	431.4	0.06	houd (hold)
in bescherming (in protection)	113.5	0.09	neem (take)
in aarde (in ground)	151.7	0.11	val (fall)
in vervulling (in fulfilment)	194.0	0.14	ga (go)
van stemming (from voting)	328.9	0.16	onthoud (refrain)
in stilzwijgen (in silence)	108.9	0.18	hul (surround)
in rekening (in account)	170.4	0.20	breng (bring)
in première (in première)	963.0	0.21	ga (go)
op adem (on breath)	170.0	0.25	kom (come)
in opstand (in revolt)	622.7	0.28	kom (come)
bij kas (in treasury)	144.3	0.29	zit (sit)
met rust (in rest)	224.3	0.35	laat (let, leave)
in verlegenheid (in embarrassment)	151.2	0.36	breng (bring)

Table 4.6: PP-Ds with low verbal entropy

(LL 6.3). On top of that, a member like *uniform* scores much higher with 78.4. We can conclude that although positive clustering results may confirm existing intuitions about certain semantic classes optionally selecting for PP-Ds, the results are not good enough for identifying those semantic classes, even if a common class exists in EuroWordNet.

4.3.6 Dependent determinerless PPs

In the previous experiments we controlled for selection by setting a minimum verbal entropy. If we focus on the other end of the scale, selecting for PP-Ds with a low verbal entropy, we find a list with a high density of phrasal verbs (see table 4.6).

Phrasal verbs are not the only verbal constructions containing PP-Ds. Some verbs take a prepositional complement which is a PP-D. In contrast to the phrasal verbs, the nominal in the PP is not fixed (22).

- (22) Van auto/baan/jurk veranderen.
of car/job/dress change
Change cars/jobs/dresses.

To find out which verb preposition combinations select for PP-Ds, we used

V	P		-D	+D
inboeten	aan	‘lose’	29	2
uitschelden	voor	‘call (so) st’	26	4
wisselen	van	‘change’	83	13
verwisselen	van	‘change’	19	4
winnen	aan	‘win’	48	12

Table 4.7: Verbs selecting for a determinerless prepositional complement

the same technique that we applied for the PP-D selecting prepositions, only changing the prepositions in verb preposition tuples: we calculated for each combination the ratio between the types with and without a determiner. The lowest ranked combinations (with proportionally the most determinerless occurrences) are listed in table 4.7. Although we excluded combinations with an uncountable noun, we still find verbs selecting for uncountable prepositional complements, such as *inboeten aan* ‘lose’, showing once again the influence of countability information on the results. As the precision of this list is low, we did not include them in further experiments.

4.4 Evaluation and Distribution of PP-Ds

We extracted a total of 363 PP-Ds from the automatically parsed corpus with various methods. Table 4.8 summarizes the PP-D types, the extraction methods and the results. To evaluate the classification methods proposed, we took all 3612 preposition+noun patterns from CGN, the syntactically annotated Corpus of Spoken Dutch Levelt (1998), removed all uncountable nouns, typos and obligatorily PP-D selecting prepositions, as well as PP-Ds containing phrases that Alpino analyzes as fixed, and classified the resulting 1510 PP-Ds according to PP-D type on the basis of our extracted data collection. From those 1510 PP-Ds, our methods classified 836 as a syntactically marked PP-D of a certain category. A total of 674 PP-Ds were *not* classified, resulting in a recall of 55.1%. We can compare this figure to a raw frequency baseline. If we take the 359 most frequent preposition+noun patterns in the training data instead of the 359 PP-Ds in our data lists, we get a recall of 34.3%. This indicates that our approach not only provides more detailed information than raw frequency (namely the PP-D type of a preposition+noun pattern), but also leads to a higher recall.

¹³43% of the extracted PPs were syntactically marked PP components of phrasal verbs. In additional 12% of the cases the PP was not syntactically marked (the noun was un-

PP-D type	N	Test	Settings	Precision (%)
Preposition	10	Type Ratio	F>50	60
Fixed PP-D	45	Cut-off	F>10, -D>99%, +P>90%, Ent<1.00%	80
Idiosyncratic PP-D	200	Log-likelihood	F>10, V Ent>2.00	63
Phrasal Verbs	108	Entropy	F>10, V Ent<2.0, LL>100	43/55 ¹³
Total	363			

Table 4.8: Extracted PP-Ds

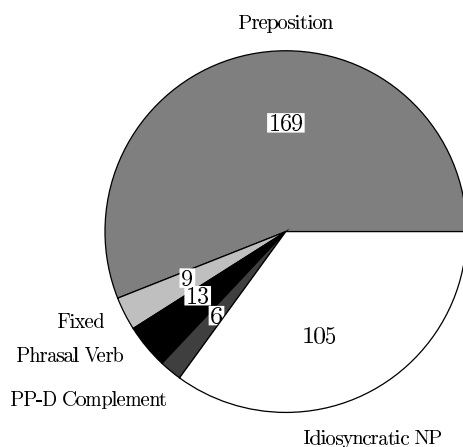


Table 4.9: Distribution of PP-D types in CGN.

The tables 4.9 and 4.10 show the results of the classification and give an impressions of the distribution of different PP-D types. We see that the prepositions that optionally select for PP-Ds and prepositions that select idiosyncratic NPs make up about 90% of the classified data. The fully productive PP-D selecting prepositions take up the largest part of the types, whereas most tokens are prepositions with idiosyncratic NPs.

Secondary evaluation on the written data from the Alpino Treebank van der Beek et al. (2002a) leads to higher recall (63.9% for semi-automatic extraction vs. 38.8% for raw frequency), but shows the same overall distribution of PP-D types.

We excluded the PP-Ds that Alpino classifies as fixed, because those PP-Ds were not represented in the training data either. This influenced the distribution in table 4.10: the 41 fixed PP-D types we excluded gave rise to 330 tokens in CGN. Among these excluded fixed PP-Ds are multiword adverbs, collocational PPs and fixed parts of larger idiomatic expressions, but by far the most tokens are phrasal verbs. The overall percentage of phrasal verbs in the data is thus higher than indicated in the pie charts.

As mentioned, 674 PP-Ds were *not* classified. These unclassified PP-Ds form a heterogeneous group. Among the negatives we find typical characteristics of spoken language (*in dinges* ‘in what’s-its-name’) and typos, but also clear PP-Ds with idiosyncratic NPs (*in bad* ‘in bath’) and many instances of the preposition *met*, seventh on the list of prepositions that optionally select for NP-Ds. Interestingly, we also find members of the clothing class (*in pyjama/smoking/uniform* ‘in pyjamas_{sg}/smoking/uniform’) and the jour-

countable), but it was part of a phrasal verb.

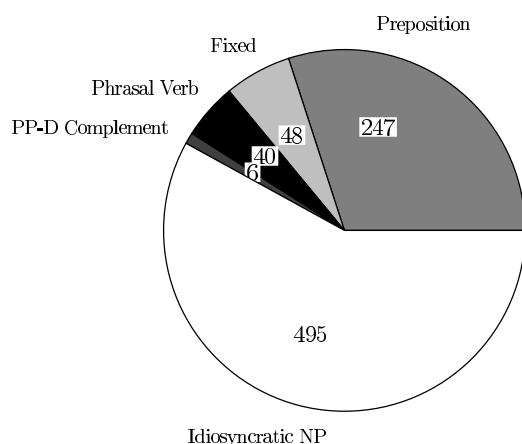


Table 4.10: Distribution of PP-D tokens in CGN.

ney class (*op tournee/trektocht/trouweis* ‘on tour/hiking tour/honeymoon’), showing that a proper treatment of semantically restricted PP-Ds has the potential of improving coverage. Another category of false negatives are compounds of words we identified as parts of PP-Ds (*in poedervorm, op beleidsniveau* ‘in powder form’, ‘at the level of policy makers’, lit. ‘on policy level’). As compounding is generally possible in all PP-Ds except the fully fixed, these false negatives will be correctly analyzed by a grammar that includes a list of nouns occurring in PP-Ds.

4.5 Conclusion and Discussion

PP-Ds form a heterogeneous collection of constructions with varying degrees of syntactic and semantic markedness. A correct analysis of the different types of PP-Ds requires knowledge about which nouns and which prepositions participate in PP-Ds, and the modifiability of the resulting structure, which is generally not available. It was illustrated that a base repository of PP-Ds can be composed semi-automatically on the basis of automatically parsed corpus data. Information about the prepositions and the nouns, their modifiers and the governing verbs was extracted and used to calculate the association between the presence or absence of a determiner and the occurrence of the noun in or outside of a PP.

Baldwin et al. (2003) and Baldwin et al. (to appear) showed that PP-Ds are not just a Dutch problem, but that they occur in many languages. The methods can be applied to other languages, as long as either a large syntactically annotated corpus is available or an unannotated corpus and a

preprocessor which can extract prepositions, nouns, verbs and modifiers from sentences with PP-Ds. The resulting repository of PP-Ds may be fed into parsing systems to extend their coverage. In the case of the Alpino parser, the PP-D repository is a first step towards abolishing the $NP \Rightarrow N$ rule, which furthermore requires more accurate countability data and high quality named entity recognizers.

The absence of gold standard data complicates thorough evaluation. Manual inspection showed that many uncountable nouns are included in the candidate lists, illustrating the importance of high quality countability information. Evaluation on CGN furthermore showed a considerable increase in recall compared to the raw frequency baseline, but with a recall of 55.3% on spoken language and 64.7% on written data, there is still plenty of room for improvement. It was shown that PP-Ds selected for by prepositions and PP-Ds composed of idiosyncratic preposition-noun combinations were the most frequent PP-D types: the two classes made up about 90% of the extracted corpus data.

In this paper, we made categorical decisions about prepositions and nouns: either they were classified as a particular type of PP-D or they were not. But virtually all preposition-noun combinations also occur *with* a determiner. Which items were included and which were excluded depended on the setting of parameters, such as N in selecting the N highest ranked PP-Ds with idiosyncratic NPs. Adjusting the parameters to include more candidates increased coverage. However, the more preposition-noun combinations are allowed to form a PP-D, the smaller the expected effect on ambiguity and the less ungrammatical PP-Ds we exclude from being parsed or generated. This trade-off between coverage and effect may be avoided if the rankings of the candidates are interpreted as weights that indicate the probability of that preposition-noun combination to participate in a PP-D.

Chapter 5

Countability

This chapter describes how the countability of nouns can be learned automatically from linguistic resources. Countability is an important lexical feature that determines the syntactic contexts in which nouns can occur, more specifically their ability to combine with or without particular determiners and quantifiers. Countability information is important for accurate and efficient parsing, generation and translation, and we also saw that it was crucial for our research on determinerless PPs in chapter 4. Unfortunately, countability information is not generally available. We therefore experiment with methods to acquire countability information automatically. Two types of linguistic resources are used in this chapter to learn countability: raw text corpora that we preprocess with various linguistic analyzers, and EuroWordNet, a lexical semantic network. In the face of sparse data, we augment our resources with English data and perform crosslingual classification.

5.1 Introduction

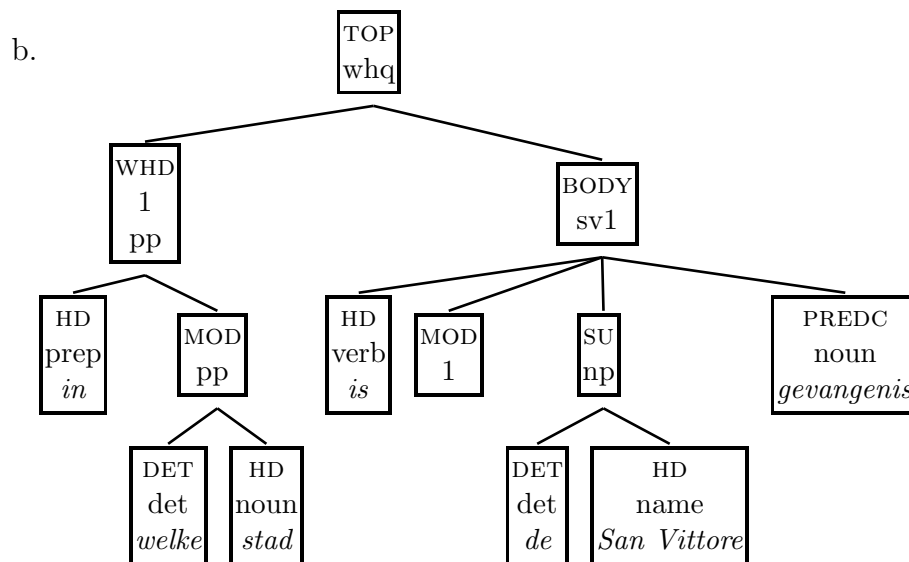
In this chapter we discuss two methods for the acquisition of lexical properties of nouns from linguistic resources. We investigate to what extent corpus data on the one hand and ontologies on the other hand are sufficient sources of information for classifying nouns according to their lexical properties, and we compare their relative performance.

We focus on the linguistic property of countability. Noun countability has not received very much attention in the computational linguistics literature (but see section 5.2 for discussion of some previous work on countability). Nevertheless, it does play an important role in (computational) grammars. The countability of a noun determines its potential to occur with (or without) particular determiners. Singular indefinites form a clear example:

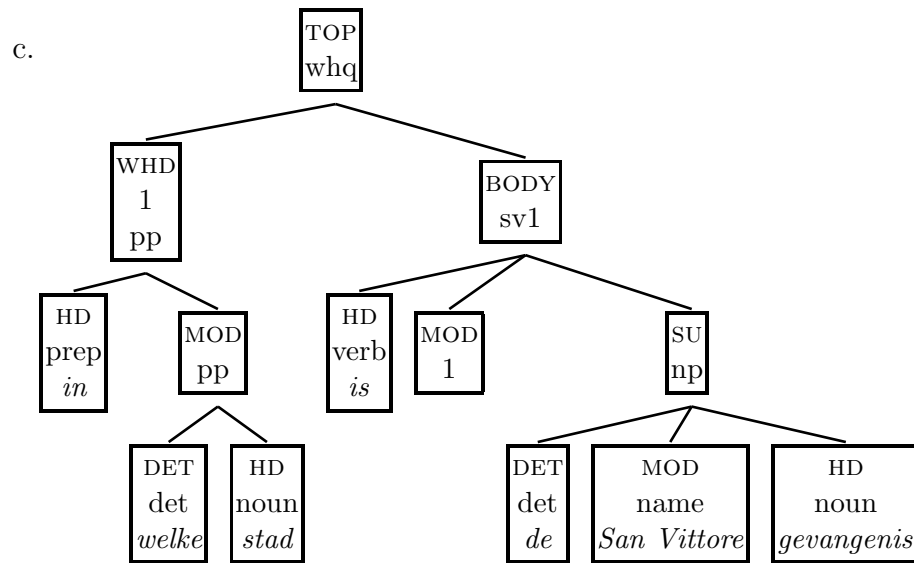
while uncountable indefinite nouns do not combine with a determiner, indefinite countable nouns obligatorily do combine with the indefinite article *een* in Dutch or *a* in English.¹

Influencing the combinatory potential of nouns, countability is important for language generation and translation. But countability information may also help to reduce the (false) ambiguity of sentences in automatic parsing. We illustrate this point with two examples. The sentences in (1) and (2) are sentences from the Alpino Treebank and the TwNC newspaper corpus. In both cases, the Alpino grammar produced both the correct parses in (1-c) and (2-c) and the false parses in (1-b) and (2-b). In the first example, the complex NP *San Vittore gevangenis* ‘San Vittore jail’ functions as the subject of the sentence, as illustrated in (1-c). However, the parser mistakenly splits the complex NP into two separate NPs: a subject NP ‘San Vittore’ and a predicative complement (PREDC) ‘jail’ (1-b). This incorrect parse was even considered the best parse. Only once the system knows that the word *gevangenis* ‘jail’ is countable in Dutch, will it correctly discard the parses in (1-b) as improbable, as the noun in itself cannot saturate an NP.

- (1) a. In welke stad is de San Vittore gevangenis?
 in which city is the San Vittore jail?
 In which city is the San Vittore jail?

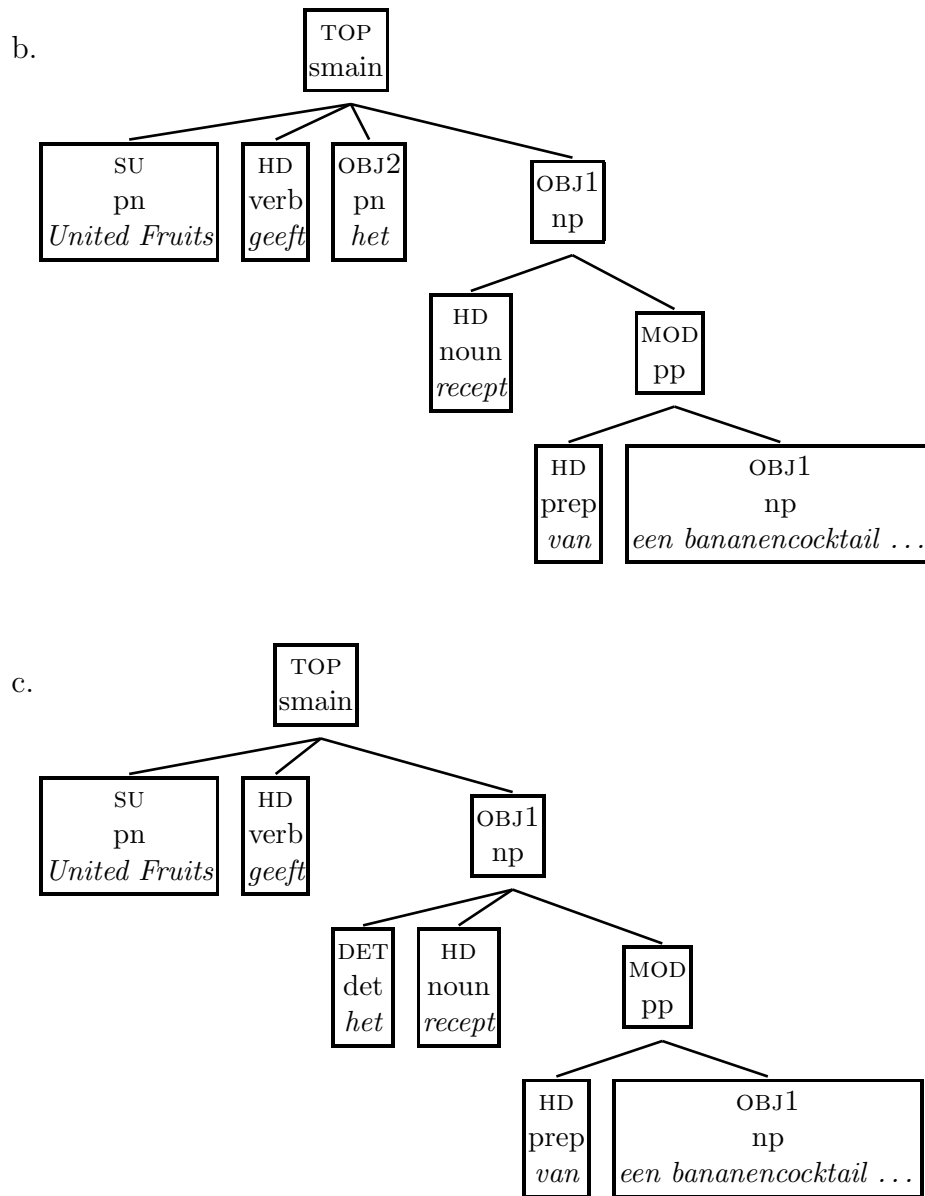


¹Except in some predicative constructions, e.g. *Ik wil matroos worden* ‘I want to be a sailor’ (lit. I want to be sailor).



Similarly, the system mistakenly splits up the complex object NP *het recept van een bananencocktail* ... ‘the recipe of a banana cocktail’ into two NPs. The word *het* is interpreted as the homonymous pronoun *het* ‘it’ (OBJ2), and analyzed as the indirect object. *Recipe of a banana cocktail* is interpreted as the direct object (OBJ1), resulting in the reading ‘United Fruits gives it recipe...’. This incorrect analysis could have been ruled out on the basis of the information that *recipe* is a count noun and thus requires a determiner. Naturally, the ambiguity reduction illustrated in (1) and (2) only works for countable nouns.

- (2) a. United Fruit [...] geeft het recept van een bananencocktail,
 United Fruits [...] gives the/it recipe of a banana cocktail
 die toepasselijk ‘Juanita’ heet.
 that appropriately Juanita is named
United Fruits gives the recipe of a banana cocktail, that is appropriately named ‘Juanita’.



Countability may also help word sense disambiguation. Often, two senses of a word have different countabilities. For example, *glas* ‘glass’ is countable in its drinking equipment sense, but uncountable in its substance reading. Based on this information, we can rule out the drinking equipment reading in (3-a) and the substance reading in (3-b). Only sentence (3-c) remains ambiguous, as the determiner *het* ‘the’ is compatible with countable and uncountable words.

- (3) a. Ik heb glas nodig.
 I have glass necessary
 I need glass.
- b. Ik heb een glas nodig.
 I have a glass necessary
 I need a glass
- c. Ik heb het glas nodig
 I have the glass necessary
 I need the glass.

Another application of countability information was described in detail in the previous chapter. It was shown that the syntactically-marked determinerless PPs can only be distinguished from the unmarked ones with accurate noun countability information.

We conclude that countability information is crucial for correct and efficient parsing and generation, as well as for the identification of certain syntactically-marked constructions, and that it can resolve certain types of (false) ambiguity. However, countability information is not generally available. In this chapter we try to fill this gap by means of automatic countability classification, experimenting with both corpus-based and ontology-based methods.

The rest of this chapter is structured as follows. In section 5.2 we discuss the notion of countability and the lexical resources that are available for English and Dutch, as well as some earlier work on countability classification. The next two sections discuss the two main approaches to countability classification that we experimented with and the results of both methods: corpus-based countability classification in section 5.3, and ontology-based classification in section 5.4. We end this chapter with a comparison of the various approaches and some concluding remarks in section 5.3.8. The research reported on in this chapter was done in close collaboration with Timothy Baldwin. Parts of this research were previously published in Baldwin and van der Beek (2003) and van der Beek and Baldwin (2004).

5.2 Preliminaries

5.2.1 Countability classes

We consider both Dutch and English to have the three countability classes of countable (also known as ‘count’), uncountable (also known as ‘mass’) and

plural only.² Countable nouns can be modified by denominators (prototypically numbers), and generally have a morphologically-marked plural form: *een fiets* ‘one bike’, *twee fietsen* ‘two bikes’. This class contains nouns which are easily individuated (i.e. there is a clear concept of a ‘base unit’ of the concept). Uncountable nouns cannot be modified by denominators, do not have a plural form, but can be modified by unspecific quantifiers such as *veel* ‘much’: **een eten* ‘one food’, *een beetje eten* ‘some food’, **twee etens* ‘two foods’. This class includes many abstract, material-denoting, collective and deverbalised nouns. Plural-only nouns have only a plural form, and cannot be enumerated: *goederen* ‘goods’, **drie goederen* ‘*three goods’. The plural-only class is considered to be a closed class in Dutch. We listed the members of this class in table 5.1. In addition to this list, there are a number of fixed expressions with plurals only nouns, e.g. example (4). As the pluralia tantum are a closed class, the classification experiments below focus exclusively on the countable and uncountable classes, ignoring nouns which are plural only.

- (4) Hij zit op zijn hurken.
 he sits on his PLURAL ONLY
 He is sitting on his heels.

It is important to realize that different senses/usages of a given word can occur with different countabilities, cf. *Ik wil een konijn* ‘I want a rabbit’ (countable) vs. *I zou graag nog wat konijn willen* ‘I would like some more rabbit, please’ (uncountable). It is not necessarily the case, however, that because a given word occurs with distinct countabilities it has multiple senses. Consider, e.g., *voor mij een rode wijn, graag* ‘for me a red wine, please’ (countable) vs. *voor mij rode wijn, graag* ‘red wine for me, please’ (uncountable), which we claim correspond to a single sense of ‘wine’.

Accounts of countability range from a purely semantically motivated feature (Jackendoff, 1991) to a completely arbitrary lexical feature in many computational grammars, including the Alpino grammar (Bouma et al., 2001). The former runs into problems when faced with different realizations of one concept in different languages, such as the Dutch *onweer* vs. English *thunderstorm*. The Dutch noun is uncountable, whereas the translation in English is countable. An account of countability in terms of a strictly arbitrary lexical feature fails to account for the semantic underpinnings and crosslingual commonalities of countability. Moreover, it implies that type-level countab-

²Haeseryn et al. (1997) use a slightly different ontology: ‘uncountable’ is used as an umbrella term for *pluralia tantum* (our plural only) and *singularia tantum* (our uncountable).

<i>bescheiden</i>	documents	<i>chemicaliën</i>	chemicals
<i>conserveren</i>	preserves	<i>contanten</i>	cash
<i>data</i>	data	<i>doeleinden</i>	purpose
<i>echtelieden</i>	marriage partners	<i>financiën</i>	finances
<i>gebroeders</i>	brothers	<i>gegevens</i>	data
<i>gelieven</i>	lovers	<i>gemoederen</i>	minds
<i>genitaliën</i>	genitals	<i>gezusters</i>	sisters
<i>goederen</i>	goods	<i>grutten</i>	rolled oats
<i>hersenen/hersens</i>	brain(s)	<i>hurken</i>	heels
<i>ingewanden</i>	intestines	<i>inkomsten</i>	incomings
<i>intimi</i>	friends	<i>kleren</i>	clothes
<i>kosten</i>	costs	<i>levensmiddelen</i>	provisions
<i>letteren</i>	literature	<i>manen</i>	mane
<i>manschappen</i>	manpower	<i>mazelen</i>	measles
<i>memoires</i>	memoirs	<i>mensenrechten</i>	human rights
<i>middeleeuwen</i>	middle ages	<i>notulen</i>	minutes
<i>omstreken</i>	surroundings	<i>ongeregelgheden</i>	riots
<i>onkosten</i>	expenses	<i>onlusten</i>	riots
<i>paperassen</i>	papers	<i>papieren</i>	official documents
<i>personalia</i>	personal data	<i>troebelen</i>	disturbances
<i>troepen</i>	troops	<i>tropen</i>	tropics
<i>waren</i>	wares	<i>waterpokken</i>	chicken-pox
<i>watten</i>	cotton-wool	<i>zemelen</i>	bran

Table 5.1: Dutch pluralia tantum.

ility distinctions are categorical, which is in fact not the case. Allan (1980) noted that prototypical countable nouns can be used in uncountable contexts, forcing a ‘substance’ interpretation (the universal grinder, e.g. *over de hele straat lag hert* ‘there was deer all over the road’) and uncountable nouns can be denumerated in certain contexts, resulting in a ‘type’ interpretation (the universal packager, e.g. *deze winkel verkoopt drie verschillende wijnen* ‘this shop sells three different wines’). This being said, nouns are generally considered to have a predominant use or basic classification as countable and/or uncountable. Copestake (1992) accounts for both the arbitrary aspects and conversion. The semantic types ‘countable’ and ‘uncountable’ are used to capture the default classification and lexical rules are provided to account for conversion from one type to the other.

Following Bond and Vatikiotis-Bateson (2002) and O’Hara et al. (2003), we assume that the countability of a noun is to a large extent predictable from its semantics. This implies that countability is generally stable across languages. But not only countability itself is stable, also the surface effects that noun countability brings about may be very similar for related language pairs. We will see that in both Dutch and English, countability influences the co-occurrence with determiners, certain prepositions and quantity denoting constructions. These two factors, the semantic grounding of countability and the similarities in the effects that countability brings about in different languages, facilitate the crosslingual approaches that we take on countability classification when faced with sparse or medium quality data problems. The crosslingual approach to countability classification taken in section 5.3 crucially relies on the grammatical similarities of countability effects in the aligned languages, while the approach in section 5.4 relies on the semantic basis of countability only.

5.2.2 Lexical resources

For Dutch, few lexical resources with countability information are available. Our Dutch training data consists solely of dictionary data extracted from the Alpino lexicon (Bouma et al., 2001). The Alpino lexicon includes all lexical information found in the Celex electronic dictionary. Countability information was first derived from the presence or absence of a plural form in Celex: all and only *singularia tantum* were considered uncountable. However, the dictionary has been extensively augmented and modified (manually) since. The total number of Dutch nouns is around 14,500. We refer to this set as Dictionary_{NL}.

In order to test the quality of the Dutch dictionary-derived data and the performance of the classifiers developed in this chapter, we manually

annotated 196 unseen Dutch nouns. The nouns were automatically extracted from the POS-tagged NRC part of the Twente Nieuws Corpus.³ It was decided that the random sample was to be a representative sample of the input of the classifier and should not be edited, leading to an occasional tag-error in the dataset. The countability judgments are based on actual usage in the Twente Nieuws Corpus. Evidence for countability class membership was extracted from the corpus automatically and checked manually. A noun was classified as a member of each class for which any valid evidence could be found, leading to a very inclusive list. For example, one grammatical example of the plural *gemakken* ‘comforts’ (e.g. *dat zijn de gemakken van het moderne leven* ‘those are the comforts of modern life’) would lead to a classification of *gemak* as countable (as well as uncountable). The complete list of nouns and the manually assigned countability judgments can be found in Appendix A. We refer to this dataset as `AnnotatedNL`.

The agreement in countability judgments between `DictionaryNL` and `AnnotatedNL` is 81.1%. Agreement figures represent the proportion of countability judgments on which both sources agree (i.e. plus or minus countable and plus or minus uncountable for each lexical item). An important part of the disagreement is caused by the fact that nouns in the Alpino dictionary are labeled either countable or uncountable, whereas the nouns in the annotated dataset are potentially labeled as both countable and uncountable.

For English, more data is available. Information about English noun countability was obtained from two lexical sources: COMLEX 3.0 (Grishman et al., 1998) and the common noun part of ALT-J/E’s Japanese-to-English semantic transfer dictionary (Bond, 2001). These two resources were combined by taking the intersection of positive and negative exemplars for each countability class. The total number of training instances is around 6,000 words; we refer to this dataset as `DictionaryEN` for the remainder of this chapter. In a similar way as for Dutch, 100 unseen nouns were hand-annotated according to actual usage in the British National Corpus (BNC: Burnard (2000)), to make up dataset `AnnotatedEN`. With this dataset, the quality of the English dictionary data was determined. We measured the agreement with `DictionaryEN` to be 85.6%, significantly higher than for the Dutch dictionary data. For an overview of the datasets, see table 5.2.

5.2.3 Past research

Past research on countability classification falls into two basic categories: corpus-based and concept-based.

³<http://wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html>

Language	Dataset	Size	Agreement (%)
English (EN)	Dictionary _{EN}	5,853	85.6
	Annotated _{EN}	98	—
Dutch (NL)	Dictionary _{NL}	14,400	81.1
	Annotated _{NL}	196	—

Table 5.2: Countability datasets

Corpus-based countability classification is based on the premise that the countability of a word type is reflected in its corpus token occurrences, in the form of co-occurrence patterns (e.g. with determiners, verbs or prepositions). Baldwin and Bond (2003a,b) applied this approach to the task of English countability classification in two forms: a) distribution-based classification and b) agreement-based classification. Distribution-based classification is based on the relative frequency of countability related features over token occurrences of a given word. For example, of all occurrences of *rabbit*, how often did it occur without a determiner, or how often did it occur in plural? Distribution-based classification thus looks for feature distribution ‘signatures’ characteristic of different countabilities. Agreement-based classification looks for convincing evidence of occurrence of one or more features which are uniquely associated with one countability class. For example, the occurrence of a singular noun with the determiner *a* is only possible for countable nouns. The output of multiple pre-processors is used to measure the degree of agreement over the occurrence of those features: an occurrence of a particular feature is used as evidence for a countability class only if multiple preprocessors have observed this fact. Noise introduced by one of the preprocessors is thus filtered out. In evaluation over the four English countability classes of countable, uncountable, plural only and bipartite using BNC data, they found distribution-based classification to be the superior method, achieving 94.6% agreement with dictionary data (or 89.2% agreement for only the countable and uncountable classes).

Schwartz (2002) also performed corpus-based countability classification, constructing an automatic countability tagger (ACT) to learn token-level noun countabilities from the BNC. The method has a coverage of around 50%, and agrees with COMLEX for 68% of the nouns marked countable and with the ALT-J/E lexicon for 88%.

In section 5.3, we attempt to apply the distribution-based methods from Baldwin and Bond (2003a,b) to Dutch. But in contrast to their work, we do not limit ourselves to monolingual classification: we perform crosslingual classification from English to Dutch as a potential solution to the problem

of sparse or low quality in-language training data.

Word-to-word countability classification uses direct lexical alignment to determine the countability of novel words from corresponding countability-annotated words. We know of no previous work that applies this strategy, but in section 5.3, we will see that when applied in a crosslingual context using English-to-Dutch word-to-word translation and transliteration data as the source of alignment, the method is remarkably accurate. Transliteration is most accurate, with an accuracy of 98.3%, but has very limited coverage.

Concept-based countability classification, as employed in 5.4, is based on the assumption that members of a given concept class or synset have the same countability. It has been applied to English by Bond and Vatikiotis-Bateson (2002) using the ALT-J/E ontology, and O'Hara et al. (2003) using the Cyc ontology and English WordNet. Bond and Vatikiotis-Bateson (2002) cite an accuracy of 78% over a 5-way classification of countability preference, whereas O'Hara et al. (2003) achieve an accuracy of 89.5% over the two-way distinction of countable/uncountable using Cyc. We are unaware of any research which has attempted the concept-based countability classification in a crosslingual context, as described in section 5.4.

5.3 Corpus-based Classification

The corpus-based approach to countability classification that we take in this section is based on the idea that a noun's countability influences the contexts in which it occurs. It influences for instance the determiners a noun combines with, but also the prepositions and measure nouns that co-occur with it. We perform supervised learning to make a feature 'signature' of each countability class. Nouns are subsequently classified as countable or uncountable based on the contexts in which they occur and which may or may not resemble the signature of a particular class. The performance of the supervised methods is then compared to unsupervised classifiers, which simply look for the occurrence of features which are uniquely associated with one particular countability class.

The supervised classification strategy heavily relies on the quality of the training data. This quality is higher for the English training data than for the Dutch data, with a difference of 4.5% accuracy. Furthermore, English and Dutch are closely related languages, which show the same surface effects of countability. For example, both languages have determiners that combine with one particular countability class, and in both languages uncountable nouns cannot be pluralized or denumerated. Given this similarity and given the fact that better training data is available for English, we

decide to experiment with crosslingual classification. The results are compared to the monolingual corpus-based results and to translation-based and transliteration-based crosslingual countability classification.

5.3.1 Feature space

Information about the contexts a noun occurs in is collected in the form of features in a feature space, following Baldwin and Bond (2003a). This feature space is made up of several feature clusters, each of which is conditioned on the occurrence of a target noun in a given construction. The features in those clusters are either one-dimensional or two-dimensional. In the first case, they are simple counts for the occurrence of the target noun in a particular context, for example with the singular determiner *een* ‘a’. In the second case, they are counts for the combination of two context factors. An example of a two-dimensional feature is the co-occurrence of the target noun in singular with the number neutral determiner *geen* ‘no’ or the co-occurrence of the target noun without a determiner and with the preposition *met* ‘with’. Below, we provide a basic description of the 9 feature clusters used in this research. After the name of the cluster, we give the number of features in the cluster, both for English (E) and for Dutch (NL). For instance, the twodimensional feature cluster subject-verb agreement is annotated ($[2 \times 2]_{\text{E}}$ vs. $[2 \times 2]_{\text{NL}}$), indicating that on both dimensions (subject number and verb number), there are two realizations possible, resulting in a total of four combinations. Each of those combinations (e.g. singular subject and singular verb) is a twodimensional feature. The value of this feature for a specific noun is the number of times it occurs in singular as the subject of a singular verb. In table 5.3 we list the predicted correlations (table based on Baldwin and Bond (2003a) and adjusted to Dutch).

Head noun number: $[2]_{\text{E}}$ vs. $[2]_{\text{NL}}$ the number of the target noun when it heads an NP. This captures the fact that countable nouns, but not uncountable nouns, have a plural form.

Subject–verb agreement: $[2 \times 2]_{\text{E}}$ vs. $[2 \times 2]_{\text{NL}}$ the number of the target noun in a subject position vs. number agreement on the governing verb. Another check for plural occurrences of a noun, indicating that it is countable.

Coordinate noun number: $[2 \times 2]_{\text{E}}$ vs. $[2 \times 2]_{\text{NL}}$ the number of the target noun vs. the number of the other noun of the conjunct. This feature is based on the assumption that while most coordinations consist of two plural or two singular conjuncts, uncountable (singular) nouns conjoin with

plural nouns more frequently than countable singulars, e.g. *hoofdpijn en tranende ogen* ‘headache and burning eyes’, *wapens en munitie* ‘arms and ammunition’. This is a gradual difference, as countable singulars are by no means impossible with plurals.

N₁ of N₂/measure noun constructions: ^{[11×2]_E vs. [11×2]_{NL} the number of the target noun (N₂) vs. the type of the N₁ in an English N₁ of N₂ construction (e.g. *a group of people*) or Dutch measure noun construction (e.g. *een groep mensen*). We have identified a total of 11 N₁ types for use in this feature cluster (e.g. COLLECTIVE, LACK, TEMPORAL). This captures the fact that measure nouns put restrictions on the countability and the number feature of their complement, e.g. *een kilo appels* ‘one kilo of apples’ vs. *een kilo suiker* ‘one kilo of sugar’.}

Occurrence in PPs: ^{[52×2]_E vs. [84×2]_{NL} the presence or absence of a determiner when the target noun occurs in singular form in a PP. The purpose of this feature is twofold: occurrence in PPs needs to be treated separate from other occurrences, because of the possibility of determinerless PPs, which does not necessarily indicate that a noun is uncountable (see chapter 4). Furthermore, some prepositions select for countable (*per* ‘per’) or uncountable (*vol* ‘full of’) complements.}

Pronoun co-occurrence: ^{[12×2]_E vs. [7×2]_{NL} what personal, reflexive and possessive pronouns occur in the same sentence as singular and plural instances of the target noun. This feature aims at capturing pronoun binding effects: uncountable nouns are not expected to bind a plural pronoun.}

Singular determiners: ^{[10]_E vs. [10]_{NL} what singular-selecting determiners occur in NPs headed by the target noun in singular form. In Dutch, these select singular count nouns (*ieder kind* ‘every child’ vs. **iedere suiker* ‘*every sugar’). In English, the determiners may also select for uncountable nouns (e.g. *much sugar*).}

Plural determiners: ^{[12]_E vs. [13]_{NL} what plural-selecting determiners occur in NPs headed by the target noun in plural form. These determiners are not expected to occur with uncountable nouns (*enkele dagen* ‘a few days’ vs. **enkele tijd* ‘*a few time’).}

Number-neutral determiners: ^{[11×2]_E vs. [13×2]_{NL} what number-neutral determiners occur in NPs headed by the target noun, and what is the number of the target noun for each. This captures the fact that these determiners combine with plural nouns if the noun is countable, but with a singular noun if it is uncountable (*minder vrije dagen* ‘less days off’ vs. *minder zout* ‘less salt’).}

Feature cluster	Countable	Uncountable
Head noun number	S, P	S
Subj-V Agreement	S, P	S
Coordinate noun number	[S,S],[P.P],[P,S]	[S,S],[S,P]
Measure nouns	[<i>een kilo</i> ‘a kilo of’,P],...	[<i>een kilo</i> ‘a kilo of’,S],...
PPs	[<i>per</i> ‘per’,S],...	[<i>vol</i> ‘full of’,S],...
Pronoun co-occurrence	[<i>hun</i> ‘them’,P],...	[<i>het</i> ‘it’,S],...
Singular determiners	[<i>ieder</i> ‘every’,S],...	-
Plural determiners	[<i>enkele</i> ‘a few’,P],...	-
Number-neutral determiners	[<i>minder</i> ‘less’,P],...	[<i>minder</i> ‘less’,S],...

Table 5.3: Predicted values for each feature cluster (S=singular, P=plural)

The Dutch and English feature clusters represent the same linguistic structures, even if the individual features are not direct translations of each other. That is, in both English and Dutch, there are determiners that select for plural (countable) nominals, and in both languages the subject and the verb agree in number. An exception is the Dutch measure noun construction (5-a). In English, the same concept (some quantity of something) is expressed with a different linguistic construction, namely with the N_1 of N_2 construction (5-b). The two bring about the same restrictions with respect to countability (5) and thus can be aligned.⁴

- (5) a. Een kilo suiker.
 a kilo sugar
 b. A kilo of sugar.
 c. *Een kilo auto.
 a kilo car
 d. *A kilo of car.

5.3.2 Methodology

We use a variety of pre-processors to map the raw data onto the types of constructions targeted in the feature clusters, namely a POS-tagger and a full-text chunker for both Dutch and English, and additionally a dependency parser for English. For Dutch, POS-tags, lemmata and chunk data were extracted from automatically generated, fully parsed Alpino output (Bouma et al., 2001). For English, we used a custom-built fnTBL-based tagger (Ngai and Florian, 2001) with the Penn tagset, morph (Minnen et al., 2001) as our lemmatiser, an fnTBL-based chunker which runs over the output of the tagger, and RASP (Briscoe and Carroll, 2002) as the dependency parser.

These data sets are then used independently to test the efficacy of the different systems at capturing features used in the classification process, or in tandem to consolidate the strengths of the individual methods and reduce system-specific idiosyncrasies in the feature values. When combining the Dutch and English in classification, we invariably combine like systems (e.g. Dutch tagger-derived data with English tagger-derived data).

The Dutch data was extracted from the newspaper (NRC, 13M words)

⁴In fact, the term ‘measure noun construction’ is an umbrella term for singular, plural and number neutral ‘measure nouns’, similar to the distinction between singular, plural and number neutral determiners and similar also to the N_1 of N_2 construction in English. Although most of these are some sort of measure, we also included nouns like *type* (*een bepaald type auto* ‘a certain type of car’)

component of the Twente Nieuws Corpus⁵ and the English data comes from the written component (90M words) of the British National Corpus (Burnard, 2000).

After generating the different feature vectors for each noun based on the above configurations, we filtered out all nouns which did not occur at least 10 times in NP head position according to the output of all pre-processors. This resulted in 20,530 English nouns and 12,734 Dutch nouns in the training data.

We propose a variety of both monolingual (Dutch-to-Dutch = NN) and crosslingual (English-to-Dutch = EN) unsupervised and supervised classifier architectures for the task of learning countability. We employ two basic classifier architectures: (1) a separate binary classifier for each countability class (**BIN**), and (2) a single multiclass classifier (**MULTI**). The multiclass classifier assigns each noun to one of the three classes ‘countable’, ‘uncountable’ or ‘both’. A classification in the category ‘both’ corresponds to a positive classification in both binary classifiers.

In all cases, our supervised classifiers are built using TiMBL version 4.2 (Daelemans et al., 2002), a memory-based classification system based on the k -nearest neighbour algorithm. TiMBL was used with the default configuration except that k was set to 9 throughout.

5.3.3 Monolingual classifiers: design

The various different monolingual classifiers determine the countability of Dutch target nouns on the basis of in-language training material. This training material consist of the 14,400 noun Alpino dictionary data (Dictionary_{NL}), for which we saw that the agreement with the hand-annotated data set was 81.1%. In this section, we discuss the binary classifiers. The multiclass classifiers (both monolingual and crosslingual) are discussed in section 5.3.7.

Unsupervised classifiers

The simplest baseline classifier simply maps all nouns to the most frequent class, which in our case is +countable and –uncountable. In addition to this, we derive a separate baseline for each countability class/pre-processor system combination. We built a (binary) monolingual unsupervised classifier based on diagnostic evidence. For each target noun, the unsupervised classifier simply checks for the existence of diagnostic data in the output of the POS tagger and chunker for the given countability class. Diagnostic data

⁵<http://wwwhome.cs.utwente.nl/~druid/TwNC/TwNC-main.html>

takes the form of unit features which are uniquely associated with a given countability class, e.g. the determiner *een* ‘a’ co-occurring with a given (singular) noun is a strong indicator of that noun being countable. We refer to these classifiers as $NN_{\text{BIN}}(\text{evidence}, \text{POS})$ and $NN_{\text{BIN}}(\text{evidence}, \text{chunk})$. We perform basic system combination by voting between the two pre-processor datasets as to whether the target noun belongs to a given countability class, and breaking ties in favour of the POS tagger ($NN(\text{evidence}, \text{all})$).

Distribution-based classifiers: $NN_{\text{BIN}}(\text{feature}_{\text{ALL}})$

We implemented a conventional monolingual classifier based on the full feature set given above (section 5.3.1). For each target noun, we compare its value for each feature with the values of other nouns on that feature and the value of the target noun on other features within the feature cluster.

As the absolute frequency of a particular feature-value combination of a noun cannot be compared with the values for other nouns or features, we follow Baldwin and Bond (2003b) in translating each one-dimensional feature f_s for target noun w into three separate feature values, representing the frequency relative to corpus size, word frequency and feature cluster frequency. By means of illustration, we calculate the relative feature values for the feature `HEAD NOUN NUMBERsg` for word w , which occurred 389 times in singular and 2 times in plural in a 10M word corpus. Suppose that 13 occurrences of the singular noun were in N-N compounds, and in all other (376+2) occurrences the noun was heading an NP. First, we capture the frequency relative to the corpus size:

$$\text{corpfreq}(f_s, w) = \frac{\text{freq}(f_s|w)}{\text{freq}(*)} \quad (5.1)$$

where $\text{freq}(*)$ is the frequency of all words in the corpus. For w , this results in: $\text{corpfreq}(\text{HNN}_{sg}, w) = 376/1,000,000 = 0.000376$. We furthermore calculate the frequency relative to the target word’s frequency:

$$\text{wordfreq}(f_s, w) = \frac{\text{freq}(f_s|w)}{\text{freq}(w)} \quad (5.2)$$

Continuing our example w , we get $\text{wordfreq}(\text{HNN}_{sg}, w) = 376/391 = 0.962$. The third relative frequency compares our count to the frequencies of the other features in the feature cluster:

$$\text{featfreq}(f_s, w) = \frac{\text{freq}(f_s|w)}{\sum_i \text{freq}(f_i|w)} \quad (5.3)$$

The third relative frequency for our example w is then $featfreq(HNN_{sg}, w) = 376/378 = 0.995$. Instead of our raw frequency of 376, we now have the 3 relative frequencies 0.000376, 0.962 and 0.995.

In addition to mapping individual unit features onto triples, we introduce a triple for each feature cluster as a whole. This triple represents the sum over all member values.

In case a feature is two-dimensional (e.g. the number of the target noun in subject-position vs. the number of the agreeing verb), each feature $f_{s,t}$ for target noun w is translated into the same relative frequencies $corpfreq(f_{s,t}, w)$, $wordfreq(f_{s,t}, w)$ and $featfreq(f_{s,t}, w)$ as above. In addition, two feature values are introduced which represent the $featfreq$ values relative to the totals of each of the two feature dimensions i and j (in combination with the target word). In other words: we calculate the frequency of target word w occurring as the singular subject of a singular verb relative to the frequency of a singular subject w with any verb (singular or plural).

$$featdimfreq_1(f_{s,t}, w) = \frac{freq(f_{s,t}|w)}{\sum_i freq(f_{i,t}|w)} \quad (5.4)$$

$$featdimfreq_2(f_{s,t}, w) = \frac{freq(f_{s,t}|w)}{\sum_j freq(f_{s,j}|w)} \quad (5.5)$$

Finally, we calculate the feature values for the cluster totals for the two-dimensional features. Where this total was a simple sum over all individual feature values for the one-dimensional feature clusters, we now calculate row and column totals. For instance, we calculate totals for each preposition (irrespective of the presence of a determiner) and for determinerless and with-determiner contexts. Each of these totals is described in the form of 3 values, similar to the individual feature values. This methodology is described in Baldwin and Bond (2003b). Given the feature space in section 5.3.1, we generate a total of 1,664 independent values for each target noun.

From the binary Alpino data, individual countable and uncountable classifiers were learned ($NN_{BIN}(feature_{ALL})$). The feature values in each case were averaged across those from the tagger and chunker.⁶

⁶We additionally built separate classifiers based on the outputs of the individual pre-processors, but found their performance to be inferior to that of the classifier based on their amalgamated output.

Classifier	Acc (%)
Baseline	74.3
NN _{BIN} (evidence,POS)	55.1
NN _{BIN} (evidence,chunk)	50.8
NN _{BIN} (evidence,all)	53.3
NN _{BIN} (feature,all)	81.9
Alpino dictionary	81.1

Table 5.4: Results for monolingual classification

5.3.4 Monolingual classifiers: results and discussion

We compare the overall performance of the different monolingual classifiers. Classifier performance is rated according to accuracy (Acc), i.e. the proportion of correct classifications (table 5.4). For each lexical item, two binary classifications are performed: it is plus or minus countable and plus or minus uncountable. Note that the classifier architecture allows for lexical items to be classified as neither countable nor uncountable.⁷ We can compare the scores relative to each other and against a simple baseline. This baseline is a majority class classifier which naively classifies all instances as belonging to the largest class (i.e. +countable and –uncountable).

The most striking result is that the unsupervised methods, which were supposed to provide an additional baseline for each combination of countability class and preprocessor, in fact perform considerably worse than our simple majority class baseline.

We zoom in on the unsupervised classification results to see if we can say more about the types of mistakes the classifiers make. We investigate the results for the classification of countable and uncountable noun separately. The results for uncountable nouns (table 5.6) are more accurate, even though the baseline for uncountable (63.8%) is much lower than for countable (84.7%). However, the accuracy remains below the baseline on both countability classes. The precision and recall scores show large differences between countable and uncountable nouns. The tables 5.5 and 5.6 show that irrespective of the preprocessing, precision (P) of the countable classifiers was high. However, this high precision was matched with a low recall (R). For the uncountable classification, we find the reverse situation: a high recall,

⁷In an engineering context, one would use the baseline classifier as a fall-back, mapping the ‘unclassified’ items to the majority class. The results presented here are based on the system as is. Interestingly, the full feature-based classifiers only failed to classify nouns that can be considered noise in the testset, caused by tag-errors.

Method	Acc (%)	P	R	F
NN _{BIN} (evidence,all)	55.1	.964	.488	.648
NN _{BIN} (evidence,chunk)	50.5	.973	.434	.600
NN _{BIN} (evidence,POS)	47.4	.970	.392	.558

Table 5.5: Unsupervised classification results for countable nouns

Method	Acc (%)	P	R	F
NN _{BIN} (evidence,all)	55.5	.423	.930	.581
NN _{BIN} (evidence,chunk)	51.0	.414	.887	.565
NN _{BIN} (evidence,POS)	63.8	.490	.718	.583

Table 5.6: Unsupervised classification results for uncountable nouns

but very low precision. In other words: the diagnostics to identify countable nouns are very accurate, but cannot be found very often. The diagnostics for uncountable nouns are more frequently found, but are not very accurate.

These findings are in line with the idea that nouns have a basic countability classification, but may be converted to another countability class: a single occurrence of a diagnostic for uncountable nouns does not mean the noun has a base classification ‘uncountable’.

These results may be improved slightly by tweaking the set of diagnostics. For example, only base prepositions were considered as diagnostics, while some collocational prepositions were also shown to pose countability restrictions on their complements in chapter 4 (e.g. *bij wijze van* ‘by means of’ selects for a countable noun). Nevertheless, the method is not expected to produce reliable lexical information, even with modifications.

The results furthermore show that the unsupervised classifiers that use POS-tagged data only outperform both the chunk-based classifier and the combination of both types of data.

The feature-based classifiers perform much better than the unsupervised classifiers: not only do they outperform the baseline, but with an accuracy of 81.9%, the corpus-based classifiers are also more accurate than the accuracy of the Alpino dictionary training data (81.1%). We expect that there is room for further improvement. As the supervised methods heavily depend on the quantity and quality of the training data, the results may be improved by training on more or better data. Unfortunately, no more or better training material is available for Dutch. For English, on the other hand, high quality dictionary data is available. Although the size of the dataset is smaller than the Dutch dataset (almost 6K words versus more than 14K for Dutch), the quality is higher, with a 85.6% agreement with the hand-annotated dataset,

versus 81.1% for Dutch. In addition to this high quality dictionary data, there are large quantities of automatically classified nouns available. Given the fact that English and Dutch are closely related languages, we decide to experiment with crosslingual classification with the higher quality English training data.

5.3.5 Crosslingual classifiers

Below, we describe two ways in which the corpus-based countability classifier can be adjusted to classify Dutch nouns based on English training data. The resulting classifier is compared to its monolingual counterpart and to two other crosslingual approaches: translation-based and transliteration-based classification. We start with a description of each of the crosslingual classifiers.

Translation-based classifier: $EN_{\text{BIN}}(\textit{translate})$

Translation-based classification applies the observation that Dutch nouns often take the same countability as their English translation equivalents. For this task we use the English automatically classified dataset, which is the output of a monolingual supervised English countability classifier (Baldwin and Bond, 2003a,b). We then extract translation pairs from a bilingual dictionary (English–Dutch freedict version 1.1-1, containing 15,426 Dutch entries) and for each countability class, vote for the membership of a given Dutch noun based on the countabilities of the English translations. In the case that no translation data exists for a given Dutch noun or no countability data exists for the English translations, we classify the Dutch noun countability as ‘unknown’. Additionally, we map plural only and bipartite nouns in English onto the Dutch uncountable class.⁸

Transliteration-based classifier: $EN_{\text{BIN}}(\textit{transliterate})$

Transliteration-based classification relies on the fact that some proportion of the Dutch nouns are spelled the same as their English translations, e.g. *tank*, *pupil*, *norm*, *item*, *restaurant*. As in translation-based classification, it applies the observation that countability is frequently preserved under translation from English to Dutch, even though some mismatches exist (e.g.

⁸This approach was chosen because the restrictions of plural only and bipartite nouns resemble those of uncountable nouns better than those of countable nouns. In hindsight, mapping of bipartite to countable may have been a better choice, as most translations of bipartite nouns are in fact countable in Dutch.

tissue, which only has a countable sense in Dutch, but has both countable and uncountable uses in English). It takes a Dutch noun and simply determines if a countability-annotated word of the same spelling exists in English, and if so, transfers the countability directly across to Dutch. In all other respects, we implement the method identically to translation-based classification. The advantage of transliteration over translation is that it is resource free. The obvious disadvantage is the expected low coverage.

Cluster-to-cluster classifier: $EN_{\text{BIN}}(\text{cluster})$

As observed above (section 5.3), there is a strong correlation between the feature clusters used for Dutch and English. For example, co-occurrence with plural determiners is a strong indicator that the given noun is countable in both English and Dutch. At the same time, there is generally low correlation between individual unit features. For example, the English plural determiner *many* has no direct Dutch equivalent, and conversely, the Dutch plural determiner *sommige* has no direct English equivalent. The most straightforward way of aligning feature clusters, therefore, is through the (three) amalgamated totals for each one-dimensional feature cluster and some subset of the column and row totals for each two-dimensional feature cluster (e.g. for the PP feature, we align the totals for the singular and plural features but not the totals for each individual preposition independent of number). All values for the individual unit features are then ignored. In this way, it is possible to align 88 feature values, based on the output of the English and Dutch POS taggers.⁹ Note that as part of the feature alignment, we take the negative log of all corpus frequency (*corpfreq*) values in an attempt to reduce the effects of differing corpus sizes in English and Dutch (about 90M words for English, vs. 13M for Dutch)

Feature-to-feature classifiers: $EN(\text{feature})$

While we stated above that there is generally low correlation between individual unit features in English and Dutch, some unit features are highly correlated crosslingually. One example is the English singular determiner *a* which correlates highly with the Dutch *een*. Here, we can thus simply match the feature values onto one another directly. In other cases, a many-to-many mapping exists between proper subsets of a given feature cluster (e.g. the

⁹All crosslingual feature-based methods were tested over the output of the POS taggers, the chunkers and the combined outputs of the three English and two Dutch pre-processors. Overall, there was very little separating the results, and the simple POS tagger generally produced the most consistent results, so it is these results we present herein.

English determiner pair *each* and *every* correlates highly with the Dutch determiner pair *ieder* and *elk*), and alignment takes the form of feature value amalgamation in each language by averaging over the unit values, followed by alignment of the amalgamated values. A total of 466 unit feature values are amalgamated into 351 feature values, which are then combined with the 88 aligned total values from cluster-to-cluster classification for a total of 439 feature values. As for cluster-to-cluster classification, we evaluate feature-to-feature classification over the output of the English and Dutch POS taggers.

We implemented a total of 5 feature-to-feature classifiers. The first, $EN_{\text{BIN}}(\text{feature}_{\text{ALL}})$, makes use of all aligned features in the form of separate binary classifiers. The second, $EN_{\text{MULTI}}(\text{feature}_{\text{ALL}})$, similarly uses all aligned features, but in a multiclass classifier architecture. The other three make use only of a subset of all features: $EN_{\text{BIN}}(\text{feature}_{\text{DET}})$ is based on only aligned determiner features, plus the aligned cluster totals; $EN_{\text{BIN}}(\text{feature}_{\text{PREP}})$ is based on only aligned preposition features, plus the aligned cluster totals; and $EN_{\text{BIN}}(\text{feature}_{\text{PRON}})$ is based on only aligned pronoun features, plus the aligned cluster totals.¹⁰

System combination: $EN_{\text{BIN}}(\text{combined})$

System combination takes the outputs of heterogeneous classifiers and makes a consolidated classification based upon them. It has been shown to be effective in tasks ranging from word sense disambiguation to tagging in consolidating the performance of component systems (Klein et al., 2002; van Halteren et al., 2001). In our case, we take the outputs of all unsupervised (i.e. evidence-based) and crosslingual classifiers—a total of 12 classifiers—for each countability class, and run TiMBL over them (effectively weighing the influence of each classifier). The 196 Dutch annotated nouns were used as training words for this procedure, and the results are thus based on 10-fold cross-validation. This provides an estimate of the classification performance we could expect over unannotated Dutch noun data using the 196 annotated nouns as training data. Finally, we experimented with a combination of the twelve classifiers above and the Alpino-trained Dutch supervised (binary) classifier.

5.3.6 Crosslingual classification: results and discussion

For a first overview of the results, we compare the accuracy (Acc) and the coverage (Cov) of the various different classifiers on the hand-annotated test

¹⁰Results for the multiclass classifier over feature subsets were found to be markedly worse than for binary classifiers.

Classifier	Acc (%)	Cov (%)
Baseline	74.3	100
NN _{BIN} (ALL)	81.9	100
EN _{BIN} (transliterate)	98.3	30
EN _{BIN} (translate)	88.6	36
EN _{BIN} (cluster _{ALL})	77.3	100
EN _{BIN} (feature _{ALL})	72.5	100
EN _{BIN} (combined)	83.2	100
E/NN _{BIN} (combined)	85.7	100

Table 5.7: Results for crosslingual classification

Feature	Acc (%)
Det	76.0
Prep	76.5
Pron	71.2
All	72.4

Table 5.8: Accuracy for various features.

set in table 5.7. The results for the baseline and the best monolingual (binary) classifiers are included for reference.

The accuracy of the simple translation and transliteration-based classifiers are surprisingly high. Their use is limited, however, because of the low recall. Where all other classifiers always give a positive or negative classification, the translation or transliteration-based classifiers can only classify if a translation or transliteration is available. Failing this, the classifier returns ‘unknown’. That is, assuming we have countability data for an English word of the same spelling as a given Dutch noun (or for a translation), we get a very accurate estimate of the Dutch countability.

The crosslingual feature and cluster-based classifiers each individually performed better than baseline, but worse than the monolingual feature-based classifier. The cluster-based classification outperforms the feature-based classification, indicating that important information is lost in the (incomplete) alignment of features. Combined with the high overhead in hand-aligning features in feature-to-feature classification, it is clear that cluster-to-cluster classification should be preferred over feature-based classification.

We investigated the effect of the various individual feature clusters. Table 5.8 list the accuracy of each cluster on the test set. The determiner cluster and the preposition cluster individually perform better than baseline, confirming the assumption that determiner and preposition co-occurrence are in-

fluenced by noun countability. However, the pronoun cluster performs worse than baseline and the combination of all clusters leads to a decrease in accuracy. We conclude that pronoun information does not contribute to the correct classification of Dutch nouns as countable or uncountable, even if it appeared helpful for English (Baldwin and Bond, 2003a). This is not to say that pronoun binding information is not influenced by noun countability, but that the proxy of this information in the present study is not a good modeling of pronoun binding information.

System combination proved helpful for countability classification. The combination of crosslingual classifiers (translation, transliteration and distribution-based) performed better than any of the component classifiers. Manually taking out the pronoun cluster (which was shown not to contribute to correct classification) did not change the results at all: running TiMBL over the outputs of all classifiers apparently (correctly) causes the pronoun cluster to receive minimal weight.

The combined crosslingual classifier outperformed monolingual classification. This confirms our claim that given the lack of reliable training data in Dutch, crosslingual classification using English data is a viable option. This finding is particularly striking given that the volume of Dutch training data is more than twice the volume of English data. Having said this, the combined crosslingual/monolingual classifier (EN/NN_{BIN}(combined)) outperforms both the combined crosslingual classifier and the monolingual classifier, in which sense the Alpino data has some empirical utility. That is, we have shown that high-quality out-of-language English countability data is a stronger predictor of Dutch countability than medium-quality in-language Dutch countability data, but at the same time that the two are complementary.

Finally, it should be noted that the classifiers perform much better for countable than for uncountable nouns. To illustrate this effect, we put together the accuracies on countable and uncountable nouns for various classifiers (see table 5.9). This is mainly due to the fact that there is a large difference in the relative occurrence of members of the two classes. Countable nouns are much more frequent than uncountable nouns, resulting in a much higher baseline for countable nouns. For illustration: 84.7% of the nouns in the gold standard data were annotated as countable, vs. 36.2% uncountable (20.9% of the nouns were annotated as having both countable and uncountable uses).

Method	Countable	Uncountable
Baseline	84.7	63.8
EN _{BIN} (translate)	94.8	58.3
EN _{BIN} (transliterate)	100.0	96.6
EN _{BIN} (cluster)	80.6	74.0
EN _{BIN} (feature _{ALL})	75.0	69.9
EN _{BIN} (combined)	86.2	79.1
EN/NN _{BIN} (combined)	88.8	78.1

Table 5.9: Accuracy for countable and uncountable nouns

5.3.7 Binary vs. three-way classifiers

For both the monolingual and the crosslingual classifiers, we presented the results of the two binary classifiers: one that classifies nouns as having or not having a countable use and one that classifies nouns as plus or minus uncountable. In addition to these binary classifiers, we implemented three-way classifiers using a selection of the classification methods used for binary classification. The three-way classifiers map the nouns to one of the three classes (strictly) countable, (strictly) uncountable or countable+uncountable. Note that the multiclass classifier does not allow the classification ‘none’, even though it is possible for a noun to receive a negative classification from both binary classifiers.

The results in table 5.10 are calculated the same way as before: the accuracy is the percentage of correct *classifications*. In other words: a classification as ‘both’ is counted as +countable, +uncountable, a classification as countable maps to +countable, –uncountable. This way, we can compare the performance of the multiclass classifier with the binary classifier. We see that the results for the threeway classifier are more stable, ranging from 80.6% to 83.4%. It performs better on the worst performing classification methods, but worse on the two best ones, so that the overall best performing classifiers are still the binary crosslingual combined classifier and the binary Dutch+English combined classifier.

5.3.8 Corpus-based approach: conclusion

We have presented several methods for classifying Dutch nouns as countable and/or uncountable on the basis of Dutch and English data. The classifiers depend on translation/transliteration data or linguistic features that were extracted from corpora. We compared a range of crosslingual English-to-

Method	Binary	Multiclass
Baseline	74.2	74.2
NL	81.9	81.4
EN cluster	77.3	80.6
EN feats	72.5	81.6
EN combined	83.2	82.7
EN/NN combined	85.7	83.4

Table 5.10: Accuracy (%) for binary and multiclass classifiers.

Dutch classifiers based on reliable English countability data with monolingual Dutch-to-Dutch classifiers based on lower-quality Dutch countability data, and found that the crosslingual classifiers outperformed the monolingual classifiers to varying degrees. Based on this, we suggest that the optimal fast-track solution to Dutch countability classification is to use English data. We were able to reach a 85.7% agreement with the hand-annotated dataset for the combined crosslingual/monolingual classifier, which is higher than the Alpino dictionary data (81.1%).

We saw that translation and transliteration-based countability classification performed remarkably well given that a translation or transliteration was available. It would be interesting to explore in more detail the possibility of co-training via translation- and transliteration-based classification, as this seems to provide a means for automatically generating high-quality Dutch countability data to learn a monolingual classifier from. The high performance of translation and transliteration-based classification furthermore supports the idea that countability is primarily connected to a semantic concept, rather than to the realization of that concept in a particular language. This hypothesis is further tested in the next section, where we classify nouns on the basis of the countability of its synonyms and other semantically related words both within a language and across languages.

5.4 Ontology-based Classification

Ontologies such as WordNet¹¹ (Fellbaum, 1998) and EuroWordNet¹² (Vossen and Bloksma, 1998) comprise a hierarchical network of concept nodes, populated with words. The nodes in such networks are conventionally termed ‘synsets’, as they contain sets of synonymous words representing a com-

¹¹<http://www.cogsci.princeton.edu/~wn/>

¹²<http://www.illc.uva.nl/EuroWordNet/>

mon underlying concept. Synsets offer a means of semantic generalization, both over the component words within a given synset and between synsets (and by extension their component words) via hierarchical relations such as hyponymy (subordination) and hypernymy (superordination). In addition, EuroWordNet connects the synsets of various languages, thus facilitating cross-linguistic generalization.

The in-language forms of generalization have been successfully applied in a variety of tasks including text categorization (e.g. de Buenaga Rodríguez et al. (2000)), PP attachment (e.g. Stetina and Nagao (1997)), subcategorization frame acquisition (e.g. Preiss et al. (2002)), selectional preference learning (e.g. Clark and Weir (2002)) and information retrieval (e.g. Mandala et al. (2000)).

This section examines the use of synsets in the automatic acquisition of the countability class of individual words. The underlying assumption is that some lexical properties are not (completely) arbitrary, but to a large extent determined by semantics, and moreover that WordNet synsets are at an appropriate level of semantic granularity to capture such properties. Under this assumption, the determination of lexical properties can be made at the synset level and applied to the individual members through simple propagation. Determination of synset-level properties is possible by inheriting the lexical properties of annotated members of a given synset. There are conflicting claims as to the semantic grounding of countability (Wierzbicka, 1988; Jackendoff, 1991; Gillon, 1996), but in terms of lexical ontologies, previous research has shown there to be a high correlation between the synset membership of English nouns and their countability classification (Bond and Vatikiotis-Bateson, 2002; O'Hara et al., 2003).

We take this line of research a step further in exploring the possibilities both for mono- and crosslingual ontology-based countability classification in English and Dutch, using EuroWordNet as our common resource. That is, we attempt to determine the countability of each synset in EuroWordNet from Dutch and/or English training data, and then evaluate the accuracy of the synset-level countability predictions over held-out data in the two languages. We thus apply the additional, cross-linguistic, generalization that EuroWordNet facilitates.

We attempt crosslingual classification, as English and Dutch are closely-related languages and the basic nature of noun countability aligns well in the two languages. We already saw in section 5.2.1 that both languages distinguish between the three countability classes of countable, uncountable and plural only,¹³ and although mismatches exist—e.g. *hersen* (plural only) vs.

¹³A fourth class of bipartite nouns (e.g. *scissors*, *trousers*) is generally recognized for

‘brain’ (countable), *onweer* (uncountable) vs. ‘thunderstorm’ (countable)—many Dutch words are in the same countability class as their English equivalents (e.g. *fiets* ‘bike’, *eten* ‘food’, *goederen* ‘goods’). Through direct comparison of monolingual and crosslingual classification, this research empirically quantifies the level of countability consistency between the two languages, relative to in-language consistency.

In the following, we first outline the lexical resources used in this research, especially where they differ from the resources used in the previous section (section 5.4.1). We then detail the classification procedures (section 5.4.2) and evaluate each method (section 5.4.3).

5.4.1 Lexical resources for WordNet-based classification

For ontology-based classification, we used all the datasets we used for the corpus-based classification, as described in section 5.2.2 and repeated in table 5.11. In addition, we used a second data set consisting of some 11,000 English nouns that were automatically classified on the basis of corpus data (Baldwin and Bond, 2003a,b). In section 5.2.3, we described the procedure that was used for the corpus-based classification. From the classified nouns, we extracted the (countable and uncountable) common nouns, which numbered about 11,000 in total; we refer to this dataset as $\text{Learned}_{\text{EN}}$. The agreement between $\text{Annotated}_{\text{EN}}$ and $\text{Learned}_{\text{EN}}$ is 82.0%, which is still slightly higher than the Dutch dictionary data. In section 5.3 from this chapter, we composed a similar dataset for Dutch. The learned Dutch countability dataset is based on combined monolingual and crosslingual corpus-based and word-to-word classification methods. The methods are applied to around 6,000 common nouns not found in the Alpino lexicon. The agreement for $\text{Learned}_{\text{NL}}$ was higher than for the Dutch dictionary data at a respectable 85.7%, almost identical to that for $\text{Dictionary}_{\text{EN}}$. As with English, we will exclusively use the combination of these datasets in evaluation, which we will refer to as **Dic+Learn_{NL}**.

Finally, we combined the English dictionary and learned datasets with the corresponding Dutch datasets to form a single multilingual dataset of about 37,000 countability-classified nouns at overall agreement of 83.1%, which we label as **Comb_{EN/NL}**. For an overview of the datasets, see table 5.11.

We used EuroWordNet to determine the synset membership of a given noun, and also to map Dutch and English synsets onto one another. Three components were used: the Dutch database of nouns, the English database

English, but has no Dutch correlate.

Language	Dataset	Size	EWN mapped	Mean EWN polysemy	Agreement (%)
English (EN)	Dictionary _{EN}	5,853	5,826	2.1	85.6
	Learned _{EN}	11,357	6,974	1.5	82.0
	Dic+Learn _{EN}	17,210	12,800	1.8	83.8
	Annotated _{EN}	98	70	1.5	—
Dutch (NL)	Dictionary _{NL}	14,400	10,407	1.9	81.1
	Learned _{NL}	5,819	2,213	1.8	85.7
	Dic+Learn _{NL}	19,661	12,088	1.9	82.4
	Annotated _{NL}	196	159	2.0	—
Dutch & English	Comb _{EN/NL}	36,871	24,888	1.8	83.1

Table 5.11: Countability datasets for WordNet-based classification

of nouns and the Inter-Lingual Index (ILI). The Dutch component contains about 35,000 nouns, grouped into synsets. The English component is a reformatted version of WordNet 1.5, and contains nearly 88,000 nouns. The ILI interconnects the monolingual ontologies by way of hyponym, hypernym, synonym and near-synonym relations. Each record in the ILI is in turn connected to the WordNet 1.5 ontology by way of one or more ‘offsets’, each representing a WordNet synset. Multiple offsets are used to collapse portions of the WordNet 1.5 structure which correspond to systematic polysemy or overly fine-grained sense distinctions, and also to add sense distinctions which are made in two or more of the languages targeted by EuroWordNet but not in the original WordNet 1.5 ontology.

In table 5.11, we present the number of nouns in each dataset which is mapped onto the EWN ontology, and also the mean polysemy of each EuroWordNet-mapped noun (i.e. the average number of senses per noun). In addition to the dictionary and the annotated data, we have added the data that was automatically learned in the previous section. Dictionary and learned data are combined in the Dic+Learn datasets. We observe that the Dic+Learn_{EN} and the Dic+Learn_{NL} datasets are very similar with respect to the number of EWN-mapped words, the agreement with the annotated datasets and the mean level of polysemy.

5.4.2 Classifier design

We experimented with classifiers that vary along two dimensions: the classification method and the EuroWordNet link types between training and test words. The classification methods we used are union-based classification,

majority-based classification and combined classification. The EuroWordNet relations we experimented with are (near-)synonyms, hypernyms, hyponyms and cohyponyms. In our first set of experiments, we test the different classification methods over (near-)synonym training words only. In a second set of experiments, we then include countability information from hypernyms and hyponyms.

While we have acknowledged that different senses of a word can occur with different countabilities, we have no immediate way of determining which EuroWordNet senses of a given word correspond to which countability.¹⁴ We are thus forced to assign the countability class(es) of each noun to all its senses in EuroWordNet.

Classification method

In this section, we detail each of the classification methods proposed in this research. We illustrate their differences by way of the Dutch noun *wederpartij* ‘antagonist/adversary’(countable) and the English-to-Dutch crosslingual classification task, using the Dictionary_{EN} dataset. In EuroWordNet, *wederpartij* maps onto WordNet offsets 6071277 (glossed as ‘a hostile person who tries to do damage to you’) and 5922580 (glossed as ‘someone who offers opposition’). English nouns mapped onto WordNet offset 6071277 are *opponent* (countable), *opposition* (uncountable) and *enemy* (countable), with the indicated countabilities in the dictionary dataset; English nouns mapped onto WordNet offset 5922580 are *adversary*, *antagonist* and *opponent*, of which the dictionary dataset lists only *opponent* as countable. In our discussion of each classification method, we discuss how this countability information is used in classifying *wederpartij*.

Union-based classification For each target noun, the union-based classifier determines the countability class(es) of all training words occurring in the synset(s) of the target noun. The noun is then assigned the union of all attested countability classes.

Under this method, *wederpartij* is classified as being both countable (by virtue of its similarity to *enemy* and *opponent*) and uncountable (by virtue of its similarity to *opposition*).

¹⁴In fact, countabilities in the ALT-J/E lexicon (Bond, 2001) are tailored to the different senses of each word, but given our partial use of its countability data and the lack of an established mapping between the ALT-J/E ontology and EuroWordNet synsets, we are unable to make use of this information.

Majority-based classification Majority-based classification is based on simple voting between the countability classes of the training words in the relevant synset(s). The target noun is assigned the (unique) most frequently attested countability class, and in the case of a tie, defaults to countable.

Under majority-based classification, *wederpartij* receives three votes for countable and one vote for uncountable, and is thus classified as being countable.

Combined classification The combined classifier maps nouns to countability classes in two steps. First, it uses majority-based classification to determine a unique classification within each synset. It then takes the union of the individual synset-based classifications. This reflects the intuition that the different countability classifications for a word are often related to the different senses of that lexical item. Also, the combined classifier is designed to filter out low-frequency countabilities in each synset a given word occurs in, hence reducing the effect of language-specific, unpredictable countability mappings of training words.

In the case of *wederpartij*, both WordNet synsets receive a countable classification, leading to the final classification of countable.

EuroWordNet link type

Synonym-based classification In synonym-based classification, we completely ignore the hierarchical structure of the ILI and use it as a simple sense inventory, expanding out each ILI record into its corresponding WordNet offset(s) (= synsets). In the crosslingual case, therefore, we end up with synsets comprising nouns in both Dutch and English.

The countability of each target noun is determined on the basis of the countability classes of those words occurring in the same WordNet synset(s), following one of the three classification methods described above.

Hypernym-based classification We also experimented with hypernym-based countability classification. The underlying (simplifying) assumption is that traversing a hypernymy link (i.e. traversing up the WordNet hierarchy) does not change the countability, and so the hypernyms can be used as additional training data in countability classification.

Classification takes place according to two steps: (1) we first look for synonyms of the target word in the training data, and if found, perform synonym-based classification; (2) failing this, we use the ILI to identify hypernym synsets of the different senses of the word, and base the class determination on training data in hypernym synsets.

Hyponym-based classification Hyponym-based classification is similar to hyponym-based classification. The only difference is that we traverse down rather than up the WordNet hierarchy via hyponym links in the second classification step, and base the countability classification on the countabilities of hyponym words.

Bidirectional classification Bidirectional classification combines hypernym and hyponym-based classification, and in the second step of classification looks both up and down the EuroWordNet hierarchy, basing classification on the combination of hypernyms and hyponyms.

We expect that the inclusion of hypernyms and hyponyms in the set of training words will lead to higher coverage (i.e. we will be able to find at least one countability for more words). On the other hand, we expect mismatches in countability to arise more frequently, e.g. *tafel* ‘table’ (countable) vs. its hypernym *meubilair* ‘furniture’ (uncountable).

Cohyponym-based classification Instead of traveling up or down the hierarchy, cohyponym-based classification looks at the countability classes of words that share a hypernym with the target word, i.e. words that are hyponyms of a noun of which the target word is a hyponym, too, a la Bond and Vatikiotis-Bateson (2002). The intuition behind this approach is that although the semantics of those sister synsets may differ considerably, the level of abstraction is the same. We thus increase the amount of training data, without introducing mismatches of the type *tafel* ‘table’ (countable) vs. its hypernym *meubilair* ‘furniture’ (uncountable), as in hypernym-based or hyponym-based classification. Instead we compare *tafel* ‘table’ with *stoel* ‘chair’, which are both countable. Similar to the other classifiers, the model is cascaded in that it only makes use of the sister information if no (countability classified) synonym is available.

5.4.3 Results and discussion

In this section we present the results for the various classification methods using each EuroWordNet link type, over different combinations of training and test datasets. We start with a basic comparison of the results for the different classification methods based on synonymy (section 5.4.3), and classify using the different EuroWordNet link types (section 5.4.3). We then present a breakdown of the results over countable and uncountable nouns (section 5.4.3), and finally contrast mono- and crosslingual classification (section 5.4.3).

Method	Accuracy (%)		Coverage (%)	
	Annotated	Dic+Learned	Annotated	Dic+Learned
Synonyms	74.8	83.3	71.8	70.4
Hyponyms	75.7	78.7	74.5	75.9
Hypernyms	74.8	79.1	97.3	97.0
Hypo+Hyper	73.4	76.1	97.3	97.4
Cohyponyms	76.4	80.6	96.6	95.5

Table 5.12: Accuracy and coverage for various link types (combined classification)

All calculations are based on test words which are contained in EuroWordNet and which have at least one countability-mapped training noun in one of the synsets accessed by the classification method in question.

Throughout evaluation, we use the combined dictionary and learned countability data for English and Dutch (i.e. Dic+Learn_{EN} and Dic+Learn_{NL}) to classify nouns in both languages. If the test set also consists of dictionary and learned data, the results are based on 10-fold cross-validation.

Performance of each EuroWordNet Link Type

The choice of a classification strategy is not independent of the choice for the link types to be included in the training data. Both vary with respect to their degree of strictness: union-based classification is more liberal than majority-based classification and hypernym-based classification is more liberal than synonym-based classification. The stricter the method, the higher the expected accuracy. But we also expect the link types to affect coverage: the more link types are included, the more training words we have and the higher the expected coverage. The classification method does not affect coverage: all methods will classify a target noun if (and only if) at least one training word is found. We therefore start by comparing the accuracy and coverage of different link types.

The results in table 5.12 confirm the intuition that the inclusion of other link types increases coverage. If we restrict ourselves to synonyms, we only find training words for about 70% of the EuroWordNet mapped nouns. Including other link types leads to a 25% increase of this proportion. The results in table 5.12 are for combined classification, but the pattern extends to other classification strategies.

It is interesting to see that the contribution of hyponym data is so much smaller than for hypernym data, even though the increase in training words

is at least as big as for hypernyms. An explanation for this fact is that while each synset in EuroWordNet (except for the top node) is associated with (at least) one hypernym synset, many nodes are terminal and thus do not have hyponyms. However, if a synset has a hyponym, it often has many. In other words, many target words have one or a few hypernyms, while few target words have many hyponyms. The effect of this distribution is that including hypernyms leads to many more target words having at least one countability mapped training word, and including hyponyms leads to a few words having many more training words than they had before (if they had any).

Looking at the effect of link type on accuracy, we see that the two data sets differ. While cross-validation on the dictionary and learned datasets shows the expected drop in accuracy, classification of the annotated test set becomes *more* accurate when including other link types. A possible explanation for this result is in the nature of the test set. It is a small test set, and the nouns were randomly selected from (POS-tagged) corpora. As a result, it contains English words (*off*, *sense*), archaic casemarkings (*state* ‘state’), nouns that are almost exclusively used in idiomatic expressions (*toom* ‘bridle’) and other non-typical nominals. That would then also explain other cases we will see later on where the results on the two test sets differ greatly. However, we would expect these not to be contained in EuroWordNet, so that they only influence coverage,¹⁵ not accuracy.

Another explanation can be found in the number of training words per target word. Compare for example the average of 4 training words per target word in synonym-based classification with the averages of 24, 22, 42 and 10 for hyponyms, hypernyms, both and cohyponyms (training data: Dic+Learn_{NL}, test data: Annotated_{NL}). While the chance of mismatches (*knife* vs. *cutlery*) increases with the introduction of more link types, classification strategies that use voting (i.e. majority-based and combined classification) may benefit from the increase in the average number of training words, as noise may be filtered out. We do not expect the larger average number of training words to increase the accuracy of union-based classification. Indeed, for union-based classification we get a synonym-based score of 76.2% on the annotated test set, vs. 74.3, 70.7 and 73.6% for hyponyms, hypernyms and cohyponyms respectively. It is unclear, however, why this effect is not found when cross-validating over the larger dictionary+learned data set.

In the remainder of this chapter, we focus on the results from cohyponym-based classification. Using hypernym, hypernym+hyponym or cohyponym relations in addition to synonym relations leads to a large increase in cov-

¹⁵That is: overall coverage, and not coverage relative to the EuroWordNet mapped nouns.

Method	Accuracy (%)	
	Annotated	Dic+Learned
Baseline	74.2	74.2
Union	73.6	73.7
Majority	75.3	81.9
Combined	76.4	80.6

Table 5.13: Cohyponym-based accuracy for different classification methods.

Classification	Union	Majority	Combined
True positive	33	4	14
True negative	57	91	84
False positive	36	2	9
False negative	18	47	37

Table 5.14: Exact cohyponym-based classification results for uncountables.

erage. The effect on accuracy varies depending on the dataset and the classification strategy, but cohyponyms consistently outperform the hypernym and hypernym+hyponym datasets with respect to accuracy.

Performance of each Classification Method

We next investigate the performance of different classification methods (table 5.13). We see that union-based classification performs below baseline. Apparently, the cohyponym data introduces noise, which the union-based classification could not filter out. This is caused by the fact that all evidence for a particular countability class directly leads to a positive classification, even if for a noun n there is only one word pointing to countability class A and a hundred words pointing to countability class B . This is also illustrated in table 5.14 which contains the exact counts for the annotated dataset: the union-based setup too easily classifies positively, resulting in high numbers of true positives, but low numbers of true negatives.

We combined classification for the remainder of the experiments. There is very little separating majority-based from combined classification, but the latter better models the intuition that countability is stable for a given synset. Furthermore, combined classification leads to the highest results overall.

Class	Baseline	Accuracy (%)	
		Annotated	Dic+Learned
Countable	84.7	84.7	82.3
Uncountable	63.8	68.1	78.9
Total	74.2	76.4	80.6

Table 5.15: Accuracy for countable and uncountable classification.

Performance over Countable and Uncountable Nouns

So far, we have averaged our results over the classification of countable nouns and uncountable nouns. However, we saw in section 5.3.6 that there were important differences between the two tasks. Most importantly, the baselines differ greatly: of our 196 item hand-annotated test set, 166 nouns were countable, whereas only 71 were uncountable (41 nouns had both countable and uncountable uses). A majority class baseline classifiers for countables thus performs with an accuracy of 84.7% while the corresponding classifier for uncountables performs with an accuracy of 63.8%.

The different baselines are reflected in the results of ontology-based classification in the same way as we saw for corpus-based classification. Table 5.15 shows the accuracy for uncountable and countable classification separately. Countable classification consistently outperforms uncountable classification. But it is striking that any gain in performance over the baseline comes from uncountable classification. In fact, a combination of a baseline classifier for countables and the cohyponym-based combined classifier for uncountables would yield the highest overall accuracy on the annotated test set, with 82%.

Mono- vs. Crosslingual Classification

The application of an ontology for countability classification was in part motivated by the fact that countability proved stable across related languages such as English and Dutch. The multilingual ontology EuroWordNet provides us with links between synsets in the Dutch network and those in the English network. This means that we can include countability mapped English words to our training set and thus expand the total amount of training data. In table 5.16, it is shown that combining Dutch and English training data leads to an increase in coverage.

More surprisingly, using English training data (with or without the Dutch data) also leads to an increase in accuracy, at least for the annotated test set. That is, crosslingual training data are a more reliable source of information

Test set	Training data					
	Dic+Learn _{NL}		Dic+Learn _{EN}		Dutch+English	
	Acc	Cov	Acc	Cov	Acc	Cov
Annotated _{NL}	76.4	96.6	80.4	92.6	80.3	98.7
Dic+Learn _{NL}	80.6	95.5	80.2	93.2		
Annotated _{EN}	82.9	58.6	75.0	51.4	81.2	68.6
Dic+Learn _{EN}	80.9	64.7	82.7	54.5		
Dic+Learn _{EN/NL}					81.4	84.3

Table 5.16: Accuracy and coverage for mono- and crosslingual classification.

than the in-language data! For the much larger dictionary+learned test set, training on English nouns exclusively leads to a 1% drop in accuracy.

We also performed the reverse classification, testing on English nouns and training on English, Dutch or combined training data. Again, we find the surprising effect that crosslingual classification is more accurate than in-language classification of the hand-annotated testset, with 82.9% for Dutch training data and only 75.0% for English training data. And again, this pattern is reversed if we test on the (English) dictionary+learned dataset. Note that the English annotated test set is even smaller than its Dutch counterpart, with only 70 nouns mapped to EuroWordNet. Finally, cross-validating over the combined Dutch+English dictionary+learned dataset gives an accuracy of 81.4%, vs. 80.3 and 81.2% on the Dutch and English annotated datasets.

We conclude with a word of caution: the annotated datasets are relatively small, and any result must therefore be interpreted with care. Having said this, 17 results are highly suggestive of the finding that using crosslingual training data has little effect on the accuracy of countability classification. Complementing in-language data with crosslingual data furthermore has a positive effect on coverage. It is therefore a successful strategy to increase the performance of classification.

Above, we discarded synonym-based classification, even if it slightly outperformed other link types with respect to accuracy, because of low coverage. But we just found that it is possible to increase coverage by complementing the training set with English data. While this gain in coverage has a modest effect on cohyponym-based classification, which has a high coverage anyway, it may improve synonym-based classification significantly. One might wonder whether synonym-based classification is still outperformed by cohyponym-based classification if more training data is available? Table 5.17 shows that on the annotated dataset, cohyponym-based classification still

Test set	Synonym		Cohyponym	
	Acc	Cov	Acc	Cov
Annotated _{NL}	79.4	91.3	80.3	98.7
Dic+Learn _{EN/NL}	83.5	76.7	81.4	84.3

Table 5.17: Accuracy and coverage of synonyms and cohyponyms on the Dutch+English dataset

has both a higher accuracy and a higher coverage. Cross-validation on the larger dataset shows a slight drop in accuracy, but much larger coverage.

If we add the majority classifier to the system as a fallback strategy, words for which no evidence can be found in the ontology will be classified +countable and –uncountable. Both synonym and cohyponym-based classifiers now have a coverage of 100%. On the annotated dataset, cohyponym-based classification outperforms synonym-based classification (79.2% vs. 78.2% accuracy), but the results for cross-validation on the dictionary+learned dataset are the other way around, with 81.1% accuracy for synonyms, and 79.8% for cohyponyms.

5.4.4 Ontology-based classification: conclusion

We have presented several methods for applying EuroWordNet to automatic countability classification, relying on the semantic grounding of countability. The proposed methods varied on two dimensions: (1) the method used to formulate a countability judgment from the training data, and (2) what links we make use of within the EuroWordNet ontology in pooling together training data. The methods were applied both to in-language and cross-language data. We showed that it is possible to learn noun countability from conceptually-linked crosslingual data, using datasets from both Dutch and English. In doing so, we demonstrated empirically that Dutch and English countabilities align as well crosslingually as they do monolingually. Combining Dutch and English data gave the best results, with an accuracy of 80.3% for the Dutch data and 81.2% for the English annotated data.

It is an interesting and yet unanswered question how this method would perform when applied to languages that are less closely related or differ with respect to the countability distinctions manifest in the languages. As the method is based only on conceptual similarity and draws its countability annotations from external sources, it can easily be applied to any language pair (assuming a common ontology and countability information in each language), even if there are divergences in the nature of countability in the two

languages.

A major drawback of ontology-based classification is the restriction that the target word must be mapped to EuroWordNet. This was the case for only 149 (76%) of the annotated nouns and 10698 (54%) of the dictionary+learned nouns. This means that even with a coverage of 100% on the EWN-mapped nouns, we will not be able to classify more than 76% and 54% of all nouns in the datasets.

5.5 Conclusion

We investigated two general methods for noun countability classification, each based on one general assumption about countability. First, we noted that a noun’s countability influences its potential to combine with particular determiners, measure nouns and so on. We used these differences to compose a ‘signature’ of syntactic contexts for each countability class, based on the corpus distribution of our training data. Target nouns were classified based on the similarity between this signature and the distribution of the target noun itself.

The second method for noun countability classification is based on the assumption that countability is stable for a given semantics, independent of its realization(s) in a particular language. EuroWordNet was applied to propagate the countability of training words to semantically related target nouns.

Both methods were used for monolingual, crosslingual and combined classification, motivated by limited in-language training data and in both cases crosslingual classification proved a viable solution to (high quality) data sparseness, performing at least as good as in-language classification. Combining mono and crosslingual classification data led to further improvements, outperforming monolingual classification and crosslingual classification independently of the classification strategy.

Of the two general approaches, the corpus-based method proved most successful. We were able to reach 85.7% accuracy on the 196 word hand-annotated test set. The ontology-based approach reached a maximum accuracy of 80.3%. On top of this, the domain of ontology-based classification is restricted to nouns that are mapped to EuroWordNet, whereas the corpus-based method can be applied to any noun occurring in the training corpus.

On the other hand, the potential for application of these methods to other language pairs, is greater for ontology-based classification than for corpus-based classification, as crosslingual corpus-based classification relies on the similarity of the two languages with respect to the surface indicators of count-

ability, whereas ontology-based classification can in principle be carried out for any two languages for which a common ontology exists.

Chapter 6

Conclusions and Future Work

In this chapter, we first summarize our main conclusions and then point to some directions for future work.

6.1 Conclusions

This thesis presented four studies in Dutch syntax. In chapter 2, we saw that the Dutch *it*-cleft construction in fact consists of *two* distinct constructions. The first is analyzed as a transitive construction with a final relative clause. The subject *het* ‘it’ and the final clause map to the same f-structure, while the focused phrase functions as the non-subject argument of the copula. The second construction is analyzed as an intransitive construction with an expletive pronoun in subject position and a final complementizer clause. A total of three constituents map to the f-structure of the subject function: the expletive, the final phrase and the focused phrase. We were thus able to formulate accounts of both constructions which conform to the rules of canonical word order without violating the principle of subject-verb agreement.

In chapter 3 it was investigated if and how the factors that are claimed to influence the English dative alternation also influence the Dutch construction. In English, the two possible realizations differ with respect to both the order of the arguments and the syntactic category of the recipient. As a result, the literature on the dative alternation includes analyses in terms of both linearization constraints and NP or PP recipients preferences. In Dutch, word order and recipient category may vary independently. We hypothesized that we would find a differentiation between on the one hand general linearization constraints influencing the order alternations in Dutch, but not the NP/PP alternation, and on the other hand construction specific constraints which influence the NP/PP alternation. This hypothesis is partially borne out.

The verb lexeme only influences the syntactic category of the recipient: it may have a preference for a PP or an NP recipient, but not for a particular order. Pronoun type and definiteness in general were shown to influence argument order. So far the results were as predicted. But the contrast between pronouns and full NPs was also shown to influence the syntactic category. And most surprising, perhaps: it was shown that the classic linearization constraint on syntactic weight (light constituents precede heavier constituents) did not influence the order of the arguments in the midfield; it only has an effect on extraposition. The influence of these linguistic factors on the distribution of the alternants is in most cases probabilistic in nature: a factor may increase the chance of finding a certain realization, but does not lead to a categorical distinction between grammatical and ungrammatical. This poses a challenge for categorical models of language.

Chapter 4 shows that the syntactically marked combination of a preposition and a bare count noun (determinerless PP or PP-D) may be the result of various different syntactic constructions. These constructions differ in productivity and modifiability. We indicated how each of these constructions could be accounted for in a grammar, given the information about which preposition and which noun may participate in a PP-D and to what extent the combination allows modification. However, this information is generally not available. It is then shown that with the help of an automatically parsed corpus and various simple statistic measures, we can extract lists of PP-Ds of particular types and their modification potential semi-automatically. The quality of this extraction and classification method heavily depends on the availability of accurate noun countability information.

Chapter 5 focuses on the automatic classification of nouns according to their countability class(es). Following earlier work on English countability (Baldwin and Bond, 2003a,b), we were able to predict a Dutch noun's base countability class from its distribution in a corpus. While the English work focused on in-language learning, we experimented also with cross-lingual classification, using English training data to classify Dutch nouns. The best results, an accuracy of 85.7%, were achieved with a combination of both mono- and cross-lingual classifiers. Translation-based and transliteration-based classification proved to be remarkably accurate, but had very restricted coverage.

The translation-based results indicate that the countability for a given semantic concept is stable across languages. Based on this observation, we made an attempt at automatic noun countability classification using the semantic ontology EuroWordNet. The nouns were classified based on the known countabilities of the target word's synonyms, hypernyms, hyperonyms and cohyponyms. Again, we experimented with both mono-lingual and

cross-lingual classification. The results for the ontology-based classification methods are not as good as for corpus-based classification, with a maximum accuracy of 80.3%.

Although the topics and the methodology in each of the chapters varied widely, corpus data was involved in all chapters. It served as a source of examples and counterexamples in our investigation in the *it*-cleft constructions of Dutch and as the source of quantitative data in our probabilistic approach to the dative construction. In chapter 4, we extracted a repository of syntactically marked PPs semi-automatically from corpus data. Finally, we developed a set of corpus-based countability classifiers, which outperformed the ontology-based classifiers significantly. We thus showed that for both theoretical linguists and computational linguists, corpora provide valuable linguistic information.

In many cases, we depended on 1) large quantities of data and 2) syntactic annotation. Although there are some useful treebanks with manually edited syntactic trees, their size is limited. We therefore decided to complement this data with corpus data that was automatically annotated with dependency trees by the Alpino parser. We were thus able to find examples of rare types of *it*-clefts in Dutch. Since the various components of *it*-clefts (*it*, *to be*, and a relative clause) are very frequent, they do not make good key words for querying raw text corpora. Searching for those frequent (function) words will give a very large set of candidate clefts, almost all of which are false positives. But syntactic annotation helps to reduce the candidate set drastically. If one can specify the syntactic relation between those frequent words, one filters out many false hits, such as simple restrictive relative clauses, while keeping the clefts in the candidate set.

The syntactic annotation also helped us to identify sentences with a double object or dative PP construction. From these sentences we obtained the quantitative data necessary to identify the linguistic factors that influence the dative alternation. Again, it would have been impossible to identify double object constructions on the basis of POS-tag sequences or word strings, and manually annotated treebanks proved too small for certain queries.

Similarly, it was the syntactic annotation which allowed us to extract information about the verb heading a determinerless PP and its modification potential in chapter 4. Although with simple POS-tags it is possible to extract a large number of the PP-Ds, it would have been hard to extract information about the verbs and the modifiers (especially the postnominal modifiers). But information about the verb is crucial for separating the verbal PP complements from the independent PP-Ds. And information about the

modifiers a PP-D co-occurs with is necessary to determine the degree in which the PP is frozen.

Chapter 5 is the only chapter in which we did not use the full annotation that the Alpino parser produces. For countability classification, we only used chunk information. However, this chunk information was extracted from the automatically generated full parses. In short: we extracted valuable linguistic information from automatically parsed corpus data in each chapter of this thesis and we applied this information to four very different types of research into Dutch syntax. The parses that were automatically generated for the corpus sentences proved to be a close approximation of their actual syntactic structure, close enough to be useful for theoretical and computational linguistic research.

6.2 Future work

The research presented in this thesis provided answers to some linguistic questions but also raised other questions that had to be left for future investigations. Why is it that pronouns have such a strong preference for the subject position, leading to the two variants of transitive clefts? What semantic feature triggers a nominal to form a complementizer cleft instead of a relative clause cleft? Why is it that weight influences extraposition, but not the position in the midfield? Will a stochastic OT implementation predict the same frequencies for the various realizations of the dative alternation as we found in the corpus? How can we best model the chance of an optionally NP-D selecting preposition to occur in a determinerless PP? Is countability best modeled as categorical with some coercion possibilities, or is it better modeled as inherently gradient? Each of these are well worth further investigation.

But above all, this thesis shows the wealth of information that has become available with the development of accurate wide-coverage parsers. Automatically annotated data facilitates research that depends on syntactic annotation. Not only for topics that are extremely common—for these topics the small manually annotated corpora may suffice—but also for less frequent phenomena. Although the automatically annotated data will no doubt contain errors, and one has to be aware of the possibility of a systematic bias in the grammar, the large difference in size (compared to manually annotated corpora) and the relatively high quality of the annotation make it a very useful resource for future research.

This thesis only scratched the surface of the possibilities. Initiatives are

being taken to apply automatically parsed data for question answering,¹ but many more applications are possible. The parsed corpora can provide quantitative data about the conditions in which for example scrambling takes place, or about ordering differences between Flemish and Dutch, and they may facilitate the study of fairly infrequent constructions such as the Dutch dative passive. Well-established analyses and new hypotheses may be tested against large quantities of data, possibly revealing exceptions that have gone unnoticed.

Corpora form a natural source of data for linguistic research, and syntactic annotation enables the linguist to extract relevant information from this source. For this, linguists no longer have to rely solely on small scale, manually-annotated corpora: they can complement this data with large, automatically annotated corpora.

¹<http://www.let.rug.nl/~gosse/Imix/>

Appendix A

ace	both	demagogie	mass
achterhoek	count	demografie	both
acrobatiek	mass	dia	count
agente	count	doel	count
ambitie	both	doodstraf	mass
arbeidersklasse	count	drang	mass
archivaris	count	dressing	both
artikelkje	count	drukke	both
asielzoeker	count	druppel	count
basketballers	count	duit	count
been	both	eb	mass
beertje	count	eeuwwisseling	count
begeleider	count	eigenaar	count
begin	count	eindpunt	count
beletsel	count	electronica	mass
bende	count	emplooi	mass
best	count	eufemisme	count
bestemming	count	exclusiviteit	both
bewaarder	count	fixatie	both
bewaker	count	foto	count
biograaf	count	freelance	mass
bluf	mass	gemak	both
brandstichting	both	gesprek	count
bravoure	mass	goed	mass
breekpunt	count	hals	count
btw	count	hamburger	both
celibaat	count	handvol	count
chanson	count	happening	count
chip	count	hechting	both
clausule	count	herwaardering	both
coma	count	hit	count
concubine	count	hoedanigheid	count
crisissituatie	count	hok	count
cylinder	count	hoofdcommissaris	count
decibel	count	hoofdschuldige	count
delcr	count	ijzerdraad	mass

inertie	mass	presentatrice	count
initiatiefnemer	count	produktiefactor	count
inkomstenderving	mass	profilering	both
item	count	programmeur	count
jungle	mass	psalm	count
kamp	count	pupil	count
kansspel	count	raadsel	count
kind	count	regent	count
klauw	count	renteverhoging	both
kneuzing	both	restaurant	count
knop	count	restitutie	both
koffertje	count	revival	count
kok	count	romanfiguur	count
koninkrijk	count	ros	count
kost	count	route	count
kringloop	count	sandaal	count
landbouwer	count	scene	count
landbouwgrond	both	scepsis	mass
leen	count	schooljaar	count
lening	both	schop	count
levensgevaar	mass	schorpioen	both
lief	count	schrikbewind	count
linksbuiten	count	sense	both
longontsteking	both	sjeik	count
maart	mass	slip	count
market	count	smash	count
mbo	count	smet	both
meligheid	mass	sneeuwstorm	count
mensheid	count	snipper	count
metaal	both	speaker	count
meting	both	spectrum	count
montage	mass	speed	mass
motiefje	count	speelzaal	count
norm	count	state	count
off	count	stimulering	both
ontreddering	mass	stop	count
opoffering	both	strateeg	count
opperbevel	mass	strijkers	count
oude	count	succes	both
ouderpaar	count	superverkiezingsjaar	count
oven	count	systematiek	both
overtreder	count	tact	mass
palm	count	tank	count
pantser	count	techno	mass
paranoia	mass	terugtocht	count
patstelling	both	tirade	count
piet	count	toom	count
poort	count	tragiek	mass
populariteit	mass	tweedeling	both

uitgedokterd	count
uitlevering	both
universiteitsbibliotheek	count
ventje	count
verbanning	both
verbreeding	both
verkrapping	both
verovering	both
verstoring	both
vertrek	count
verveling	both
verwoesting	both
verzorgingstehuis	count
vijftal	count

visserij	mass
voedingsbodem	count
voorspeller	count
wachtgelders	count
wandeling	count
wanhoop	mass
werkerrein	count
wet	count
wilg	count
zakelijkheid	both
ziel	both
zonnestraal	count

Bibliography

- Steven Abney. Statistical methods and linguistics. In Judith L. Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 1–26. The MIT Press, 1996.
- P. Adriaans, H. Fernau, C. de la Higuera, and M. van Zaanen, editors. *Special issue on grammar induction*, volume 7 of *Grammar*, 2004. Rovira i Virgili University, Tarragona, Spain.
- Alan Agresti. *Categorical Data Analysis*. John Wiley and Sons, New York, 2002.
- Judith Aissen. Markedness and subject choice in Optimality Theory. *Natural Language and Linguistic Theory*, 17(4):673–711, 1999.
- Judith Aissen. Differential object marking: Iconicity vs. economy. *Natural Language & Linguistic Theory*, 21(3):435–483, 2003.
- Adrian Akmajian. On deriving cleft sentences from pseudo-cleft sentences. *Linguistic Inquiry*, 1(2):149–169, 1970.
- Keith Allan. Nouns and countability. *Language*, 56(3):541–67, 1980.
- Joseph Aoun and Yen-hui Audrey Li. Constituency and scope. *Linguistic Inquiry*, 30:275–343, 1989.
- Jennifer E. Arnold, Thomas Wasow, Anthony Losongco, and Ryan Ginstrom. Heaviness vs. newness: The effects of complexity and information structure on constituent ordering. *Language*, 76:28–55, 2000.
- J. Atlas and S. Levinson. It-clefts, informativeness and logical form. In P. Cole, editor, *Radical Pragmatics*, pages 1–61. Academic Press, New York, 1981.

- R. H. Baayen, R. Piepenbrock, and H. van Rijn. *The CELEX Lexical Database (CD-ROM)*. Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1993.
- Timothy Baldwin, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger, and Ivan A. Sag. In search of a systematic treatment of determinerless PPs. In *Proceedings of the ACL-SIGSEM Workshop on the Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications*, Toulouse, France, 2003.
- Timothy Baldwin, John Beavers, Leonoor van der Beek, Francis Bond, Dan Flickinger, and Ivan A. Sag. In search of a systematic treatment of determinerless PPs. In *Computational Linguistics Dimensions of Syntax and Semantics of Prepositions*. Kluwer, to appear.
- Timothy Baldwin and Leonoor van der Beek. The ins and outs of Dutch noun countability classification. In *Proceedings of the 2003 Australasian Language Technology Workshop (ALTW2003)*, pages 33–40, Melbourne, Australia, 2003.
- Timothy Baldwin and Francis Bond. Learning the countability of English nouns from corpus data. In *Proceedings of the 41st Annual Meeting of the ACL*, pages 463–70, Sapporo, Japan, 2003a.
- Timothy Baldwin and Francis Bond. A plethora of methods for learning English countability. In *Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing*, pages 73–80, Sapporo, Japan, 2003b.
- S. Banerjee and T. Pedersen. The design, implementation, and use of the Ngram Statistic Package. In *Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics*, Mexico City, February 2003.
- Leonoor van der Beek. It-clefts in Dutch. Presentation at ADL01 in Paris, 2001.
- Leonoor van der Beek. The Dutch cleft constructions. In Miriam Butt and Tracy Holloway King, editors, *The Proceedings of the LFG '03 Conference*, pages 1–14. CSLI Publications, 2003.
- Leonoor van der Beek and Timothy Baldwin. Cross-lingual countability classification with EuroWordNet. In *Proceedings of the 14th Meeting of*

- Computational Linguistics in the Netherlands (CLIN 2003)*, Antwerp, Belgium, 2004.
- Leonoor van der Beek and Gerlof Bouma. The role of the lexicon in optimality theoretic syntax. In Miriam Butt and Tracy Holloway King, editors, *The Proceedings of the LFG '04 Conference*. CSLI Publications, 2004.
- Leonoor van der Beek, Gosse Bouma, Rob Malouf, and Gertjan van Noord. The Alpino Dependency Treebank. In *Computational Linguistics in the Netherlands (CLIN) 2001*, Twente University, 2002a.
- Leonoor van der Beek, Gosse Bouma, and Gertjan van Noord. Een brede computationele grammatica voor het Nederlands. *Nederlandse Taalkunde*, 7(4):353–374, 2002b.
- Otto Behaghel. Beziehungen zwischen Umfang und Reihenfolge von Satzgliedern. *Indogermanischen Forschungen*, 25:110–42, 1909/10.
- Judith Berman. On the cooccurrence of es with a finite clause in German. In T. Kiss and D. Meurers, editors, *Constraint-Based Approaches to Germanic Syntax*. CSLI, Stanford, CA, 2001.
- Rens Bod, Jennifer Hay, and Stefanie Jannedy, editors. *Probabilistic Linguistics*. MIT Press, 2003a.
- Rens Bod, Remko Scha, and Khalil Sima'an, editors. *Data Oriented Parsing*. CSLI Publications, Stanford, 2003b.
- Paul Boersma and Bruce Hayes. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32(1):45–86, 2001.
- Francis Bond. *Determiners and Number in English, contrasted with Japanese, as exemplified in Machine Translation*. PhD thesis, University of Queensland, Brisbane, Australia, 2001.
- Francis Bond and Caitlin Vatikiotis-Bateson. Using an ontology to determine English countability. In *19th International Conference on Computational Linguistics: COLING-2002*, pages 99–105, Taipei, 2002.
- P. C. Uit den Boogaart. *Woordfrequenties in geschreven en gesproken Nederlands*. Oosthoek, Scheltema & Holkema, Utrecht, 1975. Werkgroep Frequentie-onderzoek van het Nederlands.

- Kersti Börjars, Elisabet Engdahl, and Maria Andréasson. Subject and object positions in Swedish. In Mariam Butt and Tracy Holloway, editors, *Proceedings of the LFG03 Conference*, pages 43–58. CSLI Publications, 2003.
- G. Bos. Dat zijn kooplieden. *De Nieuwe Taalgids*, 54:23–27, 1961.
- Gosse Bouma and Geert Kloosterman. Querying dependency treebanks in XML. In *Proceedings of the Third international conference on Language Resources and Evaluation (LREC)*, Gran Canaria, 2002.
- Gosse Bouma, Gertjan van Noord, and Rob Malouf. Alpino: Wide coverage computational analysis of Dutch. In *Computational Linguistics in the Netherlands CLIN 2000*, 2001.
- Gosse Bouma and Begoña Villada. Corpus-based acquisition of collocational prepositional phrases. In *Computational Linguistics in the Netherlands (CLIN) 2001*, Twente University, 2002.
- Joan Bresnan. Explaining morphosyntactic competition. In Mark Baltin and Chris Collins, editors, *Handbook of Contemporary Syntactic Theory*, pages 11–44. Oxford: Blackwell Publishers, 1999. Also available on the Rutgers Optimality Archive: (ROA-299-0299).
- Joan Bresnan. Optimal syntax. In Joost Dekkers, Frank van der Leeuw, and Jeroen de Weijer, editors, *Optimality Theory: Phonology, Syntax and Acquisition*, pages 334–385. Oxford University Press, Oxford, 2000.
- Joan Bresnan. The emergence of the unmarked pronoun. In Geraldine Legendre, Jane Grimshaw, and Sten Vikner, editors, *Optimality-theoretic Syntax*, pages 113–142. The MIT Press, Cambridge, MA, 2001a.
- Joan Bresnan. *Lexical Functional Syntax*. Blackwell Publishers, 2001b.
- Joan Bresnan. The lexicon in Optimality Theory. In Suzanne Stevenson and Paola Merlo, editors, *The Lexical Basis of Syntactic Processing: Formal, Computational and Experimental Issues*, pages 39–58. John Benjamins, 2002.
- Joan Bresnan, Anna Cueni, Tatiana Nikitina, and Harald Baayen. Predicting the dative alternation. In *Proceedings of the Royal Netherlands Academy of Science Workshop on Foundations of Interpretation*, 2005. To appear.
- Joan Bresnan, Ronald. M. Kaplan, Stanley Peters, and Annie Zaenen. Cross-serial dependencies in Dutch. *Linguistic Inquiry*, 13:613–635, 1982.

- Joan Bresnan and Tatiana Nikitina. On the gradience of the dative alternation. <http://www-lfg.stanford.edu/bresnan/new-dative.pdf>, 2003.
- Ted Briscoe and John Carroll. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th Conference on Applied Natural Language Processing*, pages 356–363, Washington D.C., 1997.
- Ted Briscoe and John Carroll. Robust accurate statistical annotation of general text. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, pages 1499–1504, Las Palmas, Canary Islands, 2002.
- M. de Buenaga Rodríguez, J.M. Gómez Hidalgo, and B. Díaz Agudo. Using WordNet to complement training information in text categorization. In N. Nicolov and R. Mitkov, editors, *Recent Advances in Natural Language Processing II: Selected Papers from RANLP'97*, volume vol. 189 of *Current Issues in Linguistic Theory (CILT)*. John Benjamins: Amsterdam/Philadelphia, 2000.
- Lou Burnard. User reference guide for the British National Corpus. Technical report, Oxford University Computing Services, 2000.
- Miriam Butt, Mary Dalrymple, and Anette Frank. An architecture for linking theory in LFG. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG97 Conference*, University of California, San Diego, 1997. CSLI Publications.
- Anna Cardinaletti and Michal Starke. Deficient pronouns: a view from germanic. In Höskuldur Thráinsson, Samuel David Epstein, and Steve Peter, editors, *Studies in Comparative Germanic Syntax, Volume II*, pages 21–65. Kluwer Academic Publishers, 1996.
- Hye-Won Choi. *Optimizing Structure in Context: Scrambling and Information Structure*. PhD thesis, Stanford University, 1996.
- Noam Chomsky. Quine’s emperical assumptions. In D. Davidson and J. Hintikka, editors, *Words and objections: Essays on the work of W. V. Quine*, pages 53–68. Dordrecht: Reidel, 1969.
- Noam Chomsky. Deep structure, surface structure and semantic interpretation. In *Studies on Semantics and Generative Grammar*. Mouton, the Hague edition, 1972.
- Noam Chomsky. *Barriers*. The MIT Press, Cambridge, MA, 1986.

- S. Clark and D. Weir. Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206, 2002.
- Peter Collins. The indirect object construction in English: an informational approach. *Linguistics*, 33:35–49, 1995.
- Anne Copestake. *Lexical rules in constraint-based grammar*. PhD thesis, University of Sussex, Brighton, 1992.
- P.A. Coppen. Als dit 't nou eens was. Verschenen in NEDER-L, sept 1996, 1996. <http://www.neder-l.nl/archieven/miniaturtjes>.
- Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. TiMBL: Tilburg memory based learner, version 4.2, reference guide. Technical Report 02-01, ILK, 2002.
- M. Dalrymple, R.M. Kaplan, III Maxwell, J.T., and A. Zaenen, editors. *Formal Issues in Lexical-Functional Grammar*. CSLI Publications, Stanford University, 1995.
- Mary Dalrymple. *Lexical Functional Grammar*. Academic Press, 2001.
- Renaat Declerck. *Studies on copular sentences, clefts and pseudoclefts*. Leuven University Press/Foris Publications, 1988.
- Gerald Patrick Delahunty. *Topics in the syntax and semantics of English Cleft Sentences*. PhD thesis, University of California, 1981.
- Paul Diderichsen. *Elementaer dansk grammatik*. Copenhagen: Gyldendal, 1946.
- J.E. Emonds. *A Transformational Approach to English Syntax*. Academic Press, New York, 1976.
- N. Erteschik-Shir. Discourse constraints on dative movement. In T. Givon, editor, *Syntax and Semantics 12: Discourse and Syntax*. New York: Academic Press, 1979.
- Yehuda Falk. *Lexical-functional grammar: an introduction to parallel constraint-based syntax*. CSLI Publications, 2001.
- Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

- Michal Finkelstein-Landau and Emmanuel Morin. Extracting semantic relationships between terms: Supervised vs. unsupervised methods. In *Proc. of the Workshop on Ontological Engineering on the Global Information Infrastructure (at EKAW'99)*, 1999.
- Annette Frank, Tracy Holloway King, Jonas Kuhn, and John T. Maxwell III. Optimality theory style constraint ranking in large-scale LFG grammars. In Peter Sells, editor, *Formal and Empirical Issues in Optimality Theoretic Syntax*, pages 367–397. CSLI Publishers, Stanford University, 2001.
- G. Geerts and H. Heestermans, editors. *Van Dale Groot Woordenboek der Nederlandse Taal*. Van Dale Lexicografie, Utrecht, Antwerp, 1992.
- B. Gillon. The lexical semantics of English count and mass nouns. In *Proc. of the ACL-SIGLEX Workshop on the Breadth and Depth of Semantic Lexicons*, pages 51–61, Santa Cruz, USA, 1996.
- Talmy Givón. Direct object and dative shifting: Semantic and pragmatic case. In Frans Plank, editor, *Objects. Towards a Theory of Grammatical Relations*, pages 151–182. New York: Academic Press, 1984.
- Adele Goldberg. *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago: University of Chicago Press, 1995.
- Georgia M. Green. *Semantics and Syntactic Regularity*. Indiana University Press, 1974.
- Ralph Grishman, Catherine Macloed, and Adam Myers. *COMLEX Syntax Reference Manual*. NYU, 1998. Prometheus Project.
- Mila Groot. *Lexiconopbouw: microstructuur*, 2000. Intern rapport van het project Corpus Gesproken Nederlands.
- A. Gulli and A. Signorini. The indexable web is more than 11.5 billion pages. In *Proceedings of WWW2005, Poster Session*, pages 902–903, Chiba, Japan, 2005. Available at <http://www2005.org/cdrom/content2.htm>.
- Jeanette K. Gundel. *On the source of 'it' in cleft sentences*. Indiana University Linguistics Club, 1976.
- Jeanette K. Gundel. Universals of topic-comment structure. In Michael Hammond, Edith A. Moravcsik, and Jessica R. Wirth, editors, *Studies in syntactic typology*, pages 209–39. Amsterdam: John Benjamins, 1988.

- W. Haeseryn et al., editors. *Algemene Nederlandse Spraakkunst*. Nijhoff, Groningen, 1997.
- Hans van Halteren, Jakub Zavrel, and Walter Daelemans. Improving accuracy in word class tagging through combination of machine learning systems. *Computational Linguistics*, 27(2):199–230, 2001.
- Jack Hawkins. *A Performance Theory of Order and Constituency*. Cambridge University Press, 1994.
- Marti Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING92*, Nante, France, 1992.
- Nancy Hedberg. The referential status of clefts. *Language*, 76(4):891–920, 2000.
- Rodney Huddleston and Geoffrey K. Pullum, editors. *The Cambridge grammar of the English language*. Cambridge University Press, 2002.
- R. S. Jackendoff. \bar{X} *Syntax: A Study of Phrase Structure*. The MIT Press, Cambridge, MA, 1977.
- Ray Jackendoff. Parts and boundaries. In Beth Levin and Steven Pinker, editors, *Lexical and Conceptual Semantics*, pages 1–45. Blackwell Publishers, Cambridge, MA & Oxford, UK, 1991.
- Otto Jespersen. *A Modern English Grammar III*. George Allen and Unwin, 1927.
- Mark Johnson. Optimality-theoretic Lexical Functional Grammar. In Paola Merlo and Suzanne Stevenson, editors, *The Lexical Basis of Sentence Processing: Formal, computational and experimental issues*, pages 59–73. John Benjamins Publishing Company, Amsterdam, 2002.
- Dan Jurafsky. Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In Rens Bod, Jennifer Hay, and Stefanie Jannedy, editors, *Probabilistic Linguistics*, chapter 3, pages 39–96. MIT Press, 2003.
- R.M. Kaplan and J. Bresnan. Lexical Functional Grammar: A formal system for grammatical representation. In J. Bresnan, editor, *The mental Representation of Grammatical Relations*, pages 173–281. The MIT Press, Cambridge, MA, 1982. Reprinted in Dalrymple et al. (1995, pp. 29-130).
- R.M. Kaplan, J.T. Maxwell III, and A. Zaenen. Functional uncertainty. CSLI Monthly Newsletter, Stanford University, 1987.

- R.M. Kaplan and A. Zaenen. Long-distance dependencies, constituent structure and functional uncertainty. In M. Baltin and A. Kroch, editors, *Alternative Conceptions of Phrase Structure*, pages 17–42. Chicago University Press, 1989. Reprinted in Dalrymple et al. (1995, pp. 137–165).
- R.M. Kaplan and A. Zaenen. West-Germanic verb clusters in LFG. In Pieter A.M. Sueren and Gerard Kempen, editors, *Verb Constructions in German and Dutch*, pages 127–150. John Benjamins, 2003.
- Katia Kermanides, Nikos Fakotakis, and George Kokkinakis. Automatic acquisition of verb subcategorization information by exploiting minimal linguistic resources. *International Journal of Corpus Linguistics*, 9(1):1–28, 2004.
- Dan Klein, Kristina Toutanova, H. Tolga Ilhan, Sepandar D. Kamvar, and Christopher D. Manning. Combining heterogeneous classifiers for word-sense disambiguation. In *Proc. of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, Pittsburgh, USA, 2002.
- Manfred Krifka. Lexical representations and the nature of the dative alternation. Presentation at the University of Amsterdam. Available on <http://amor.rz.hu-berlin.de/~h2816i3x/>, November 9 2001.
- Jonas Kuhn. Corpus-based learning in stochastic OT-LFG - experiments with a bidirectional bootstrapping approach. In Miriam Butt and Tracy Holloway King, editors, *Proceedings of the LFG '02 Conference*, National Technical University of Athens, Athens, 2002. CSLI Publications.
- Jonas Kuhn. *Optimality-Theoretic Syntax—A Declarative Approach*. CSLI Publications, Stanford, CA, 2003.
- Maria Lapata. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th Meeting of the North American Chapter of the Association*, pages 397–404, College Park, MD, 1999.
- S. Lapointe. *A Theory of Grammatical Agreement*. PhD thesis, University of Massachusetts at Amherst, 1980.
- Richard K. Larson. On the double object construction. *Linguistic Inquiry*, 21:589–632, 1988.

- Hanjung Lee. Prominence mismatch and markedness reduction in word order. *Natural Language & Linguistic Theory*, 21(3):617–680, 2003.
- Jürgen Lenerz. Zur Syntax der Pronomina im Deutschen. *Sprache und Pragmatik*, 29, 1992.
- Pim Levelt. Corpus gesproken Nederlands, 1998. zie ook www.elis.rug.ac.be/cgn/.
- Beth Levin. *English Verb Classes and Alternations. A Preliminary Investigation*. The MIT Press, Cambridge, Massachusetts, 1993.
- Beth Levin and Malka Rappaport Hovav. What alternates in the dative alternation? Presentation at the 2002 Conference on Role and Reference Grammar: New Topics in Functional Linguistics: The Cognitive and Discursive Dimension of Morphology, Syntax and Semantics, Universidad de La Rioja, Logrono, Spain, 2002. Available on <http://www-csli.stanford.edu/~beth/pubs.html>.
- D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL98*, Montreal, Canada, 1998.
- Robert Malouf and Gertjan van Noord. Wide coverage parsing with stochastic attribute value grammars. In *IJCNLP-04 Workshop Beyond Shallow Analyses - Formalisms and statistical modeling for deep analyses*, 2004.
- R. Mandala, T. Tokunaga, and H. Tanaka. Query expansion using heterogeneous thesauri. *Information Processing and Management*, 36(3):361–78, 2000.
- Christopher D. Manning. Probabilistic syntax. In Rens Bod, Jannifer Hay, and Stefanie Jannedy, editors, *Probabilistic Linguistics*, chapter 8, pages 289–341. The MIT Press, 2003.
- Jason Merchant. 'Pseudosluicing': Elliptical clefts in Japanese and English. In A. Alexiadou et al., editors, *ZAS Working Papers in Linguistics*. Zentrum für Allgemeine Sprachwissenschaft: Berlin, 1998.
- P.J. Merckens. Zijn dat kooplieden of zijn kooplieden dat? *De Nieuwe Taalgids*, 54:152–154, 1961.
- Walt Detmar Meurers. On the use of electronic corpora for theoretical linguistics. case studies from the syntax of german. *Lingua*, 2004. <http://ling.osu.edu/~dm/papers/meurers-03.html>.

- Guido Minnen, John Carroll, and Darren Pearce. Applied morphological processing of English. *Natural Language Engineering*, 7(3):207–23, 2001.
- Michael Moortgat, Ineke Schuurman, and Ton van der Wouden. CGN syntactische annotatie. Internal report corpus Gesproken Nederlands., 2001.
- Gereon Müller. Harmonic alignment and the hierarchy of pronouns in german. In Horst Simon and Heike Wiese, editors, *Pronouns - Grammar and Representation*, pages 205–232. Amsterdam: Benjamins, 2001.
- Grace Ngai and Radu Florian. Transformation-based learning in the fast lane. In *Proc. of the 2nd Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL2001)*, pages 40–7, Pittsburgh, USA, 2001.
- Gertjan van Noord, Gosse Bouma, Rob Koeling, and Mark-Jan Nederhof. Robust grammatical analysis for spoken dialogue systems. *Journal of Natural Language Engineering*, 5(1):45–93, 1999.
- T. O’Hara, N. Salay, M. Witbrock, D. Sneider, B. Aldag, S. Bertolo, K. Panton, F. Lehmann, M. Smith, D. Baxter, J. Curtis, and P. Wagner. Inducing criteria for mass noun lexical mappings using the Cyc KB and its extensions to WordNet. In *Proceedings of the Fifth International Workshop on Computational Semantics (IWCS-5)*, Tilburg, The Netherlands, 2003.
- Steven Pinker. *Learnability and Cognition. The Acquisition of Argument Structure*. The MIT Press, Cambridge, Massachusetts, 1989.
- Jessie Pinkham and Jorge Hankamer. Deep and shallow clefts. In R.E. Grossman, L.J. San, and T.J. Vance, editors, *Papers from the Eleventh Regional Meeting of the Chicago Linguistic Society*, pages 429–450, 1975.
- Lonneke van der Plas and Gosse Bouma. Syntactic contexts for finding semantically related words. In *Proceedings of the 15th Meeting of Computational Linguistics in the Netherlands (CLIN 2005)*, Leiden, the Netherlands, 2005.
- Carl Pollard and Ivan Sag. *Head-driven Phrase Structure Grammar*. University of Chicago / CSLI, 1994.
- J. Preiss, Korhonen, A., and T. Briscoe. Subcategorization acquisition as an evaluation method for WSD. In *Proc. of the 3rd International Conference on Language Resources and Evaluation (LREC 2002)*, Las Palmas, Canary Islands, 2002.

- A. Prince and P. Smolensky. *Optimality theory: constraint interaction in generative grammar*. Rutgers University Center for Cognitive Science, Piscataway, NY, 1993.
- Ellen Prince. A comparison of wh-clefts and *it*-clefts in discourse. *Language*, 54:883–906, 1978.
- Ellen Prince. The ZPG letter: Subjects, definiteness, and information status. In Sandra Thompsons and William Mann, editors, *Discourse Description: Diverse analyses of a fundraising text*, pages 295–325. Amsterdam: John Benjamins, 1992.
- Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, and Jan Svartvik. *A comprehensive grammar of the English language*. Longman, 1985.
- Tanya Reinhart. Interface economy: Focus and markedness. In C. Wilder, H.M. Geartner, and M. Bierwisch, editors, *The role of economy principles in Linguistic Theory*. Akademik Verlag, Berlin, 1996.
- Luigi Rizzi. The fine structure of the left periphery. In L. Haegeman, editor, *Elements of Grammar: Handbook in Generative Syntax*. Kluwer Academic Publishers, Dordrecht, 1997.
- J. de Rooy. Onzijdige voornaamwoorden en het naamwoordelijk gezegde. *De nieuwe taalgids*, 63:181–186, 1970.
- John Robert Ross. Nouniness. In Osamu Fujimura, editor, *Three Dimensions of Linguistic Theory*, pages 137–257. TEC Company, 1973.
- Hotze Rullman and Jan-Wouter Zwart. On saying *Dat*. In Roel jonkers, Edith Kaan, and Anko Wiegel, editors, *Language and Cognition 5*, Groningen, 1996.
- Ivan Sag. English relative clause constructions. *Journal of Linguistics*, 33 (2):431–484, 1997.
- Ivan Sag and Thomas Wasow. *Syntactic Theory; A Formal Introduction*. CSLI Publications, 1999.
- L.O. Schwartz. Corpus-based acquisition of head noun countability features. Master's thesis, Cambridge University, Cambridge, UK, 2002.
- M. Silverstein. Hierarchy of features and ergativity. In R.M.W. Dixon, editor, *Grammatical categories in Australian languages*, pages 227–244. Australian Institute of Aboriginal Studies, Canberra, 1976.

- R.J.C. Smits. *Eurogrammar; The Relative and Cleft Constructions of the Germanic and Romanic Languages*. Foris, Dordrecht, 1989.
- Paul Smolensky. On the internal structure of the constraint component Con of UG. online, 1995. ROA-86-0000, <http://roa.rutgers.edu>.
- Paul Smolensky and Géraldine Legendre. *The Harmonic Mind: From Neural Computation To Optimality-Theoretic Grammar Vol. 1: Cognitive Architecture; vol. 2: Linguistic and Philosophical Implications*. MIT Press, 2005.
- Jan Philipp Soehn and Manfred Sailer. At first blush on tenterhooks. about selectional restrictions imposed on non-heads. In Gerhard Jeager, Paola Monachesi, and Gerald Penn and Shuly Winter, editors, *Proceedings of Formal Grammar 2003*, pages 149–161, 2003.
- J. Stetina and M. Nagao. Corpus based PP attachment ambiguity resolution with a semantic dictionary. In *Proc. of the 5th Annual Workshop on Very Large Corpora*, pages 66–80, Hong Kong, 1997.
- Laurel Smith Stvan. *The Semantics and Pragmatics of Bare Singular Noun Phrases*. PhD thesis, Northwestern University, 1998.
- Benedikt M. Szmrecsányi. On operationalizing syntactic complexity. In G. Purnelle, C. Fairon, and A. Dister, editors, *Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis*, volume II, pages 1032–1039, Louvain-la-Neuve, March 10-12, 2004, 2004. Presses universitaires de Louvain.
- Höskuldur Thráinsson. Object shift and scrambling. In Mark Baltin and Chris Collins, editors, *The Handbook of Contemporary Syntactic Theory*, pages 148–202. Oxford: Blackwell, 2001.
- Hans Uszkoreit. *Word order and constituent structure in German*. CSLI, 1987.
- Theo Vennemann. Explanation in syntax. In J. Kimball, editor, *Syntax and Semantics II*. New York: Academic Press, 1973.
- M. Begoña Villada Moirón. Distinguishing prepositional complements from fixed arguments. In *Proceedings of 11th EURALEX International Congress*, volume III, pages 935–942, Lorient, France, 2004.
- Piek Vossen and Laura Bloksma. Categories and classifications in EuroWordNet. In *Proceedings of First International Conference on Language Resources and Evaluation*, Granada, 1998. URL www.let.uva.nl/~ewn/.

- M. de Vries et al., editors. *Woordenboek der Nederlandsche taal*. Nijhoff [etc.], 1882-1998.
- Thomas Wasow. *Postverbal Behavior*. CSLI Publications, 2002.
- Anna Wierzbicka. *The Semantics of Grammar*. John Benjamin, 1988.
- Annie Zaenen and Ronald M. Kaplan. Formal devices for linguistic generalizations: West Germanic word order in LFG. In Mary Dalrymple, Ronald M. Kaplan, John T. Maxwell III, and Annie Zaenen, editors, *Formal Issues in Lexical-Functional Grammar*, pages 215–239. CSLI Publications, 1995.
- C. Jan-Wouter Zwart. Clitics, scrambling, and head movement in Dutch. In *Approaching Second: Second Position Clitics and Related Phenomena*, pages 579–612. CSLI Publications, 1996.
- C. Jan-Wouter Zwart. *Morphosyntax of Verb Movement; A Minimalist Approach to the Syntax of Dutch*. Dordrecht: Kluwer, 1997.

Samenvatting

Een corpus is 1) een verzameling documenten of 2) [taalk.] een begrensde verzameling teksten voor linguïstisch onderzoek (Van Dale online woordenboek). Het onderzoek in dit proefschrift richt zich op die tweede betekenis van corpora. In tegenstelling tot een corpus is taal zelf onbegrensd: met een eindig aantal bouwblokken kunnen oneindig veel taaluitingen gemaakt worden. Sommige taalkundigen hebben dan ook hun twijfels over het nut van corpora voor taalkundig onderzoek. Dit proefschrift laat zien dat corpora, ondanks de beperkingen die zij hebben, toch veel bij kunnen dragen aan allerlei soorten taalkundig onderzoek, variërend van theoretische taalkunde tot het automatisch leren van de lexicale eigenschappen van woorden.

Corpora zijn tegenwoordig veelal elektronisch. Dat maakt het mogelijk om met behulp van computerprogramma's interessante taalkundige informatie uit de bestandenverzameling te halen. En hoewel corpora per definitie begrensd zijn, neemt de omvang van de beschikbare elektronische corpora nog steeds toe. Eén jaargang krantentekst is al snel zo'n 17 miljoen woorden. En dan is er nog het web, met een geschatte 11,5 miljard pagina's eind januari 2005 (Gulli and Signorini, 2005) het ultieme corpus.

Het nut van digitale corpora voor taalkundig onderzoek kan nog vergroot worden door de tekst te verrijken met taalkundige meta-informatie. Zo kunnen de woorden voorzien worden van woordsoortlabels (*part-of-speech-tags*), kunnen de grenzen van woordgroepen toegevoegd worden (*chunks*) en kunnen de grammaticale relaties tussen die verschillende woordgroepen aangeduid worden (*parsing*). Deze verrijking van een corpus maakt het mogelijk om naar abstracte taalkundige patronen te zoeken. Zo kunnen bijvoorbeeld passieve zinnen uit een corpus gehaald worden door te zoeken naar zinnen waarin het onderwerp van *worden* overeenkomt met het lijdend voorwerp van het hoofdwerkwoord. Helaas is het handmatig annoteren van tekst heel tijdrovend. Handgeannoteerde corpora zijn dan ook beperkt in omvang. Maar voor onderzoek naar sommige (infrequente) constructies zijn juist heel grote corpora nodig. In dat geval kan een automatisch geannoteerd corpus uitkomst bieden. Hoewel automatische annotatie fouten bevat, blijkt

de meta-informatie toch van nut: in dit proefschrift worden vier uiteenlopende onderwerpen uit de grammatica van het Nederlands behandeld, waarbij automatisch verrijkte corpora telkens een andere rol spelen.

Het eerste onderwerp betreft de gekloofde zin. Dit zijn zinnen zoals in (1)–(2), die gebruikt worden om een bepaald zinsdeel te benadrukken (het meest benadrukte woord is in hoofdletters gedrukt). Ze bestaan uit het voornaamwoord *het* (soms *dit* of *dat*), het werkwoord *zijn*, de benadrukte woordgroep en een ondergeschikte bijzin. De meeste analyses van dit type zinnen—dat overigens in heel veel talen voorkomt—gaan ervan uit dat de zinnen (1) en (2) voorbeelden van een en dezelfde constructie zijn. Aangetoond wordt dat dit in ieder geval voor het Nederlands niet het geval is.

- (1) Het is immers niet de TRAINER die kansen voor open doel verknalt.
- (2) Het was op ZIJN aandringen, dat ik de redactie van de adviesaanvraag [...] zo heb veranderd.

Enkele argumenten voor het onderscheid tussen deze twee typen: in zinnen zoals (1) is de bijzin altijd een relatieve bijzin, terwijl het in zinnen zoals (2) een onderschikkende bijzin is met het voegwoord *dat*. Het benadrukte zinsdeel verschilt ook: de eerste constructie benadrukt alleen NP's, terwijl de tweede constructie allerlei woordgroepen kan benadrukken. Verder is er een verschil in het voornaamwoord in de beide zinnen. In het zinnen van het eerste type is het niet expletief, en in zinnen van het tweede type wel. Dat blijkt onder meer uit het feit dat de eerste in het corpus ook voorkomt met een demonstratief voornaamwoord *dat* of *dit* in plaats van *het*, de tweede type niet (en geconstrueerde voorbeelden bleken ongrammaticaal). Tenslotte kunnen in zinnen zoals (1) ook andere koppelwerkwoorden dan *zijn* voorkomen, maar in het type (2) niet.

Gekloofde zinnen van het type (1) worden geanalyseerd als koppelwerkwoordzinnen. Het onderwerp *het* en de relatieve bijzin vormen het onderwerp, en de benadrukte woordgroep is het predikaat. Het type (2) daarentegen heeft slechts één syntactisch argument: het onderwerp, bestaande uit de bijzin plus de benadrukte woorgroep (*het* is in dit geval semantisch leeg). Deze analyse verklaart de verschillen tussen de beide typen en de schijnbare incongruentie tussen het onderwerp en de persoonsvorm én is conform de algemeen veronderstelde regels voor woordvolgorde in het Nederlands, in tegenstelling tot sommige eerdere analyses. De analyse wordt geformaliseerd binnen het theoretisch kader van *Lexical Functional Grammar* (LFG).

In dit hoofdstuk leveren corpora voorbeelden (om de eigen analyse te onderbouwen en de taalkundige eigenschappen van de constructie te illustreren) en tegenvoorbeelden (om de tekortkomingen van alternatieve analyses aan

te tonen). Omdat de constructie laagfrequent is, moet een groot corpus gebruikt worden, en omdat de constructie alleen aan de grammaticale rollen van de woordgroepen te herkennen is, moet een syntactisch geannoteerd corpus gebruikt worden. Om deze redenen zijn automatisch geannoteerde corpora gebruikt in aanvulling op handmatig geannoteerde corpora.

Het tweede onderwerp is de meewerkend-voorwerpconstructie. Het meewerkend voorwerp kan in het Nederlands, net als in veel andere talen, gerealiseerd worden als een zelfstandig-naamwoordgroep (NP) (3) of als een voorzetselgroep (PP) (4).

(3) Heeft hij je dat niet verteld?

(4) Als de speaker die treffer abusievelijk aan Amokachi toekent, grijpt hulptrainer Jo Bonfrère in.

Er zijn 2 typen analyses van deze constructie in het Engels. Het eerste formuleert voorkeuren voor bepaalde woordgroepsoorten (“werkwoorden die een manier van communiceren uitdrukken krijgen een prepositioneel meewerkend voorwerp”), het tweede maakt gebruik van algemene orderingsprincipes (“korte zinsdelen voor lange”). In combinatie met de strikte Engelse woordvolgorde, die dicteert dat een naamwoordelijk meewerkend voorwerp vóór het lijdend voorwerp komt, maar een PP erna, leiden die algemene principes tot de keuze voor een NP (korte meewerkende voorwerpen) of een PP (lange meewerkende voorwerpen). In het Nederlands is de woordvolgorde minder strikt dan in het Engels, en hoeft de PP niet altijd achteraan te staan. De hypothese is dan ook dat algemene orderingsprincipes geen invloed hebben op de keuze voor NP of PP, maar alleen op de volgorde van de beide complementen. De corpusdata bevestigen deze hypothese op sommige punten, maar laten een niet voorspelde invloed van pronominaliteit op de woordgroepsoort zien. Bovendien wordt aangetoond dat gewicht, tegen de verwachting in, de volgorde van de argumenten in het middenveld niet beïnvloedt. Wél hebben zware voorwerpen een voorkeur voor extrapositie.

Hoewel de regels voor de meewerkend-voorwerpconstructie vaak als categoriaal worden gezien, blijken veel contrasten niet zwart-wit te zijn. Zo hebben veel werkwoorden een voorkeur voor een PP of een NP, maar zelden is het alternatief echt onmogelijk. Voor onderzoek naar dergelijke verschijnselen is corpusmateriaal onontbeerlijk. Met behulp van corpora kunnen verschillen in de frequenties van twee constructies gemeten worden, en kunnen bovendien contextfactoren geïdentificeerd worden die op deze frequentie van invloed zijn. Om deze invloeden vervolgens te modeleren, zijn de gebruikelijke taalmodellen op basis van absolute regels of constraints niet toereikend. Optimality Theory, dat een stochastische implementatie kent, is hiervoor beter

geschikt. Met een vaste ordening van constraints voorspelt het model de meest frequente varianten. Bovendien kunnen door middel van herschikking van de constraints ook de minder frequente realisaties voorspeld worden. Verder onderzoek moet aantonen of een stochastische implementatie dezelfde frequenties voorspelt als aangetroffen in het corpus. Duidelijk is in elk geval dat frequentie-informatie onontbeerlijk is in de analyse van de meewerkend-voorwerpconstructie.

Een derde verschijnsel dat onder de loep genomen wordt is de voorzetselgroep zonder determinator (PP-D). Enkelvoudige telbare woorden komen in het algemeen niet voor zonder determinator (bijv. een lidwoord of een telwoord): **Ik koop huis* is ongrammaticaal, net als **huis is mooi*. Opvallend genoeg kan dat vaak wél binnen een voorzetselgroep: *ik ga naar huis*. Geïllustreerd wordt hoe de eigenschappen van verschillende typen PP-D's in grammatica-regels (LFG) gedefinieerd kunnen worden. Maar om de goede PP's (*naar huis*) van de slechte (**naar auto*) te kunnen onderscheiden moet ook bekend zijn welke zelfstandige naamwoorden en welke voorzetsels in zo'n constructie kunnen voorkomen. Bovendien moet bekend zijn wat de grammaticale eigenschappen van die specifieke combinatie zijn, bijvoorbeeld of een bijvoeglijk naamwoord is toegestaan (**naar hoog huis* vs. *op hoge leeftijd*). Met behulp van grote corpora en eenvoudige statistische toetsen is het mogelijk om PP-D's (semi-)automatisch te identificeren en te classificeren naar hun grammaticale eigenschappen.

We onderscheiden 3 basistypen PP-D met elk hun eigen grammaticale eigenschappen. Vaste verbindingen, zoals *in zwang* of *van lieverlee* hebben een betekenis die niet volgt uit de betekenis van de afzonderlijke delen. Ze zijn in corpora te herkennen doordat het zelfstandig naamwoord niet (meer) voorkomt buiten deze constructie. Zo komt het woord *lieverlee* alleen maar voor in combinatie met *van*. Een tweede type is de compositionele PP-D. De betekenis van de PP is wél regelmatig af te leiden en vaak kunnen er ook (bepaalde) bijvoeglijk naamwoorden in voorkomen, bijvoorbeeld *in (wankel) evenwicht*. Kenmerkend is dat de combinatie van voorzetsel en zelfstandig naamwoord vaker zonder determinator voorkomt dan op basis van kans verwacht zou worden. Dit wordt met behulp van de statistische toets *log-likelihood ratio* gemeten. Het derde basistype PP-D wordt gevormd met een voorzetsel uit een kleine groep voorzetsels die verplicht (*per*) of optioneel (bijv. *zonder*) combineren met een zelfstandig-naamwoordgroep zonder determinator. Hoe meer verschillende PP-D's een prepositie vormt, hoe sterker de voorkeur voor deze combinatie. Naast deze driedeling moet ook nog onderscheid gemaakt worden tussen 'zelfstandige' PP-D's en voorzetselgroepen die alleen zonder determinator voorkomen in combinatie met een bepaald werkwoord, bijv. *in toom houden* of *van auto veranderen*. Deze twee meta-

categorieën kunnen onderscheiden worden door een minimum aan variatie in werkwoorden vast te stellen, gemeten door middel van de statistische toets *entropy*.

We kunnen nu voorzetselgroepen classificeren door in een corpus na te gaan of het de karakteristieke eigenschappen vertoont van een bepaald type PP-D. De meeste van de kenmerken zijn echter alleen te herkennen in een syntactisch geannoteerd corpus. Bovendien zijn de statistische toetsen alleen betrouwbaar bij grote hoeveelheden data. Daarom wordt opnieuw automatisch geannoteerde data gebruikt. Met behulp van deze data wordt automatisch een verzameling PP-D's samengesteld. Handmatige evaluatie toont aan dat 20-50% van de geëxtraheerde PP-D's niet syntactisch gemarkeerd is, omdat het zelfstandig naamwoord ontelbaar is en dus geen determinator behoeft. Betere informatie over telbaarheid zou dan ook leiden tot een hogere precisie. Zolang deze niet beschikbaar is, blijft handmatige evaluatie onmisbaar, maar levert extractie op basis van automatisch geannoteerde data een goede kandidatenlijst.

Voor een nauwkeurige extractie van syntactisch gemarkeerde PP-D's is het essentieel dat nauwkeurige informatie over de telbaarheid van zelfstandige naamwoorden beschikbaar is. En niet alleen daarvoor: een sprekende computer moet weten of het **ik wil tosti* is of *ik wil een tosti*. En wanneer hij de zin *ik heb een glas nodig* hoort of leest, dan moet hij weten dat het gaat om een object waaruit gedronken kan worden, niet om een bouw materiaal, zoals in *ik heb glas nodig*. Helaas is deze informatie niet in ruime mate beschikbaar. Maar met behulp van—alweer—automatisch geannoteerde corpora is het mogelijk om de telbaarheid van woorden automatisch te achterhalen met een hogere precisie dan voorheen voor handen was.

Telbare woorden verschillen van niet-telbare woorden in de context waarin ze voorkomen: telbare woorden komen voor in meervoud, niet-telbare niet (pluralia tantum buiten beschouwing gelaten); enkelvoudige telbare woorden hebben vrijwel altijd een determinator bij zich, niet-telbare niet noodzakelijkerwijs; telbare woorden kunnen voorkomen met het lidwoord *een*, niet-telbare niet. Door de contexten te bekijken van woorden waarvan we zeker weten dat ze telbaar of niet-telbaar zijn, kan een profiel gemaakt worden van de distributie van telbare en niet-telbare woorden. Wanneer nu de distributie van het testwoord in het corpus genoeg lijkt op het profiel van telbare woorden, wordt het als 'telbaar' geclassificeerd. Wanneer de distributie genoeg lijkt op die van de ontelbare woorden, wordt het woord als 'ontelbaar' geclassificeerd. Het kan voorkomen dat een woord zowel telbaar als ontelbaar is, bijvoorbeeld het woord *vis*: het dier is telbaar, het voedsel is ontelbaar. De classificatie vindt plaats op basis van *Memory-Based Learning*, een techniek voor automatisch leren.

De kwaliteit van de classificatie hangt samen met de hoeveelheid trainingsdata (woorden waarvan de telbaarheid bekend is) en voor het Nederlands is er maar weining van die data beschikbaar. Maar omdat het Nederlands en het Engels wat betreft telbaarheid erg op elkaar lijken (de effecten van telbaarheid in corpora zijn ongeveer hetzelfde), gebruiken we niet alleen Nederlandse data, maar ook Engelse: in dat geval wordt het profiel bepaald op basis van Engelse trainingswoorden, en worden daarmee Nederlandse testwoorden geclassificeerd. De resultaten van deze automatische classificatie komen voor 85,7% overeen met die van de handmatige classificatie van een testset, wat een verbetering is ten opzichte van eerder werk op dit gebied.

Er zijn ook andere manieren om de telbaarheid van een zelfstandig naamwoord te bepalen. Zo is betoogd dat de telbaarheid van een woord geen arbitraire lexicale eigenschap is, maar direct samenhangt met de betekenis van een woord. In dat geval zouden het Engelse en het Nederlandse woord voor een bepaald begrip dezelfde telbaarheid moeten hebben, net als het Franse en het Spaanse. De praktijk leert dat dit in veel gevallen ook het geval is. Op basis van deze observatie is een tweede classificatiemethode ontwikkeld. Deze methode maakt gebruik van EuroWordNet. In dit semantische netwerk zijn woorden die een bepaald concept verwoorden (synoniemen, bijvoorbeeld *meel* en *bloem*) gegroepeerd in *synsets*. Deze synsets hebben verbindingen met woorden uit verschillende talen. De synset van *meel* is dus ook verbonden met het Engelse *flour*. Daarnaast is het net hiërarchisch geordend, zodat de hypernymen *meel* en *bloem* direct boven de hyponym *tarwebloem* staan. Wanneer nu een trainingswoord in dezelfde synset voorkomt als een testwoord, kunnen we ervan uitgaan dat ze dezelfde telbaarheid hebben. We geven dus de categorie van het trainingswoord door aan het testwoord. De telbaarheid kan niet alleen doorgegeven worden aan synoniemen, maar ook aan hypo- of hyper- of cohyponymen ('zusjes' in de hiërarchie), en zowel aan Nederlandse als anderstalige (Engelse) trainingswoorden kunnen worden gebruikt. Hoewel de resultaten van dit classificatiesysteem beter zijn dan van een simpel baselinesysteem, halen ze het niet bij de resultaten op basis van corpusdata.

De vier hier samengevatte hoofdstukken illustreren heel verschillende toepassingen van corpusdata: het vinden van voorbeelden en tegenvoorbeelden, het verkrijgen van kwantitatieve data, de extractie van syntactisch gemarkeerde constructies en het automatisch leren van lexicale eigenschappen. In al deze gevallen waren grote hoeveelheden data nodig, die bovendien voorzien moesten zijn van taalkundige annotatie. Hoewel er ruimte blijft voor verbetering, leidde het gebruik van automatisch geparseerde data—ondanks de ruis door mogelijke parseerfouten—in alle gevallen tot goed bruikbare resultaten.

Groningen dissertations in linguistics

GRODIL

1. Henriëtte de Swart (1991). *Adverbs of Quantification: A Generalized Quantifier Approach.*
2. Eric Hoekstra (1991). *Licensing Conditions on Phrase Structure.*
3. Dicky Gilbers (1992). *Phonological Networks. A Theory of Segment Representation.*
4. Helen de Hoop (1992). *Case Configuration and Noun Phrase Interpretation.*
5. Gosse Bouma (1993). *Nonmonotonicity and Categorical Unification Grammar.*
6. Peter I. Blok (1993). *The Interpretation of Focus.*
7. Roelien Bastiaanse (1993). *Studies in Aphasia.*
8. Bert Bos (1993). *Rapid User Interface Development with the Script Language Gist.*
9. Wim Kosmeijer (1993). *Barriers and Licensing.*
10. Jan-Wouter Zwart (1993). *Dutch Syntax: A Minimalist Approach.*
11. Mark Kas (1993). *Essays on Boolean Functions and Negative Polarity.*
12. Ton van der Wouden (1994). *Negative Contexts.*
13. Joop Houtman (1994). *Coordination and Constituency: A Study in Categorical Grammar.*

14. Petra Hendriks (1995). *Comparatives and Categorical Grammar*.
15. Maarten de Wind (1995). *Inversion in French*.
16. Jelly Julia de Jong (1996). *The Case of Bound Pronouns in Peripheral Romance*.
17. Sjoukje van der Wal (1996). *Negative Polarity Items and Negation: Tandem Acquisition*.
18. Anastasia Giannakidou (1997). *The Landscape of Polarity Items*.
19. Karen Lattewitz (1997). *Adjacency in Dutch and German*.
20. Edith Kaan (1997). *Processing Subject-Object Ambiguities in Dutch*.
21. Henny Klein (1997). *Adverbs of Degree in Dutch*.
22. Leonie Bosveld-de Smet (1998). *On Mass and Plural Quantification: The case of French 'des'/'du'-NPs*.
23. Rita Landeweerd (1998). *Discourse semantics of perspective and temporal structure*.
24. Mettina Veenstra (1998). *Formalizing the Minimalist Program*.
25. Roel Jonkers (1998). *Comprehension and Production of Verbs in aphasic Speakers*.
26. Erik Tjong Kim Sang (1998). *Machine Learning of Phonotactics*.
27. Paulien Rijkhoek (1998). *On Degree Phrases and Result Clauses*.
28. Jan de Jong (1999). *Specific Language Impairment in Dutch: Inflectional Morphology and Argument Structure*.
29. H. Wee (1999). *Definite Focus*.
30. E-H. Lee (2000). *Dynamic and Stative Information in Temporal Reasoning: Korean tense and aspect in discourse*.
31. Ivilin P. Stoianov (2001). *Connectionist Lexical Processing*.
32. Klarien van der Linde (2001). *Sonority substitutions*.
33. Monique Lamers (2001). *Sentence processing: using syntactic, semantic, and thematic information*.

34. Shalom Zuckerman (2001). *The Acquisition of "Optimal" Movement.*
35. Rob Koeling (2001). *Dialogue-Based Disambiguation: Using Dialogue Status to Improve Speech Understanding*
36. Esther Ruigendijk (2002). *Case assignment in agrammatism: a cross-linguistic study*
37. Anthony J. Mullen (2002). *An Investigation into Compositional Features and Feature Merging for Maximum Entropy-Based Parse Selection.*
38. Nanette Bienfait (2002). *Grammatica-onderwijs aan allochtone jongeren.*
39. Dirk-Bart den Ouden (2002). *Phonology in Aphasia: Syllables and Segments in Level-Specified Deficits.*
40. Rienk Withaar (2002). *The Role of the Phonological Loop in Sentence Comprehension.*
41. Kim Sauter (2002). *Transfer and Access to Universal Grammar in Adult Second Language Acquisition.*
42. Laura Sabourin (2003). *Grammatical Gender and Second Language Processing: An ERP Study.*
43. Hein van Scie (2003). *Visual Semantics.*
44. Lilia Schürcks-Grozeva (2003). *Binding and Bulgarian.*
45. Stasinos Konstantopoulos (2003). *Using ILP to Learn Local Linguistic Structures.*
46. Wilbert Heeringa (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance.*
47. Wouter Jansen (2004). *Laryngeal Contrast and Phonetic Voicing: A Laboratory Phonology Approach to English, Hungarian and Dutch.*
48. Judith Rispens (2004). *Syntactic and Phonological Processing in Developmental Dyslexia.*
49. Danielle Bougairé (2004). *L'approche communicative des campagnes de sensibilisation en santé publique au Burkina Faso: les cas de la planification familiale, du sida et de l'excision.*

50. Tanja Gaustad (2004). *Linguistic Knowledge and Word Sense Disambiguation*.
51. Susanne Schoof (2004). *An HPSG Account of Nonfinite Verbal Complements in Latin*.
52. Begoña Villada (2005). *Data-driven identification of fixed expressions and their modifiability*.
53. Robbert Prins (2005). *Finite State Pre-Processing for Natural Language Analysis*.
54. Leonoor van der Beek (2005). *Topics in Corpus-Based Dutch Syntax*.

GRODIL

Secretary of the Department of General Linguistics

Postbus 716

9700 AS Groningen

The Netherlands